A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts

Cataldo Musto, Giovanni Semeraro, Marco Polignano

Department of Computer Science
University of Bari Aldo Moro, Italy
{cataldo.musto,giovanni.semeraro,marco.polignano}@uniba.it

Abstract. The exponential growth of available online information provides computer scientists with many new challenges and opportunities. A recent trend is to analyze people feelings, opinions and orientation about facts and brands: this is done by exploiting Sentiment Analysis techniques, whose goal is to classify the polarity of a piece of text according to the opinion of the writer.
In this paper we propose a lexicon-based approach for sentiment classification of Twitter posts. Our approach is based on the exploitation of widespread lexical resources such as SentiWordNet, WordNet-Affect, MPQA and SenticNet. In the experimental session the effectiveness of the approach was evaluated against two state-of-the-art datasets. Preliminary results provide interesting outcomes and pave the way for future research in the area.

Keywords: Sentiment Analysis, Opinion Mining, Semantics, Lexicons

1 Background and Related Work

Thanks to the exponential growth of available online information many new challenges and opportunities arise for computer scientists. A recent trend is to analyze people feelings, opinions and orientation about facts and brands: this is done by exploiting Sentiment Analysis [13, 8] techniques, whose goal is to classify the polarity of a piece of text according to the opinion of the writer.

State of the art approaches for sentiment analysis are broadly classified in two categories: *supervised approaches* [6, 12] learn a classification model on the ground of a set of labeled data, while *unsupervised* (or *lexicon-based*) ones [18, 4] infer the sentiment conveyed by a piece of text on the ground of the polarity of the word (or the phrases) which compose it. Even if recent work in the area showed that supervised approaches tend to overcome unsupervised ones (see the recent SemEval 2013 and 2014 challenges [10, 15]), the latter have the advantage of avoiding the hard-working step of labeling training data.

However, these techniques rely on (external) lexical resources which are concerned with mapping words to a categorical *(positive, negative, neutral)* or numerical sentiment score, which is used by the algorithm to obtain the overall

sentiment conveyed by the text. Clearly, the effectiveness of the whole approach strongly depends on the goodness of the lexical resource it relies on. As a consequence, in this work we investigated the effectiveness of some widespread available lexical resources in the task of sentiment classification of microblog posts.

2 State-of-the-art Resources for Lexicon-based Sentiment Analysis

SentiWordNet: SentiWordNet [1] is a lexical resource devised to support Sentiment Analysis applications. It provides an annotation based on three numerical sentiment scores (positivity, negativity, neutrality) for each WordNet synset [9]. Clearly, given that this lexical resource provides a synset-based sentiment representation, different senses of the same term may have different sentiment scores. As shown in Figure 1, the term *terrible* is provided with two different sentiment associations. In this case, SentiWordNet needs to be coupled with a Word Sense Disambiguation (WSD) algorithm to identify the most promising meaning.

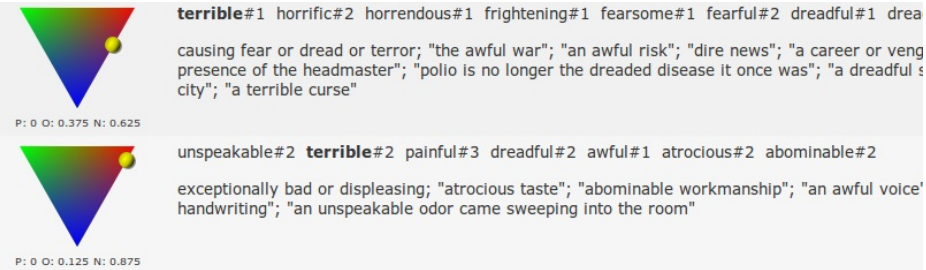

Fig. 1. An example of sentiment association in SentiWordNet

WordNet-Affect: WordNet-Affect [17] is a linguistic resource for a lexical representation of affective knowledge. It is an extension of WordNet which labels affective-related synsets with affective concepts defined as A-Labels (e.g. the term *euphoria* is labeled with the concept *positive-emotion*, the noun *illness* is labeled with *physical state*, and so on). The mapping is performed on the ground of a domain-independent hierarchy (a fragment is provided in Figure 2) of affective labels automatically built relying on WordNet relationships.

MPQA: MPQA Subjectivity Lexicon [19] provides a lexicon of 8,222 terms (labeled as *subjective expressions*), gathered from several sources. This lexicon contains a list of words, along with their POS-tagging, labeled with polarity (positive, negative, neutral) and intensity (strong, weak).

SenticNet: SenticNet [3] is a lexical resource for *concept-level* sentiment analysis. It relyies on the Sentic Computing [2], a novel multi-disciplinary paradigm for Sentiment Anaylsis. Differently from the previously mentioned resources, SenticNet is able to associate polarity and affective information also to complex

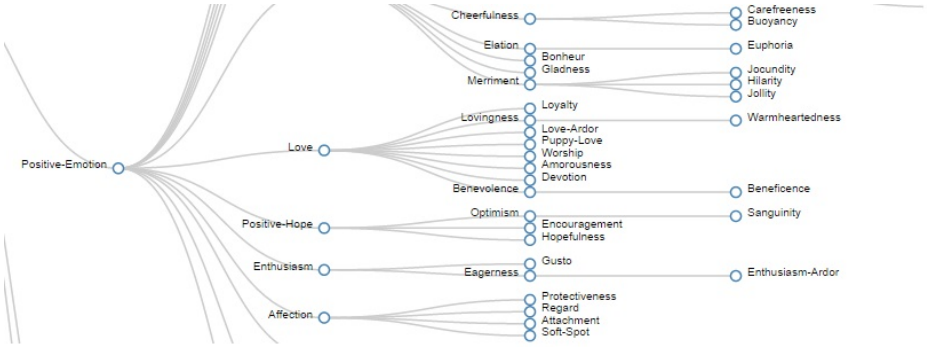

Fig. 2. A fragment of WordNet-Affect hierarchy

concepts such as *accomplishing goal*, *celebrate special occasion* and so on. At present, SenticNet provides sentiment scores (in a range between -1 and 1) for 14,000 common sense concepts. The sentiment conveyed by each term is defined on the ground of the intensity of sixteen *basic* emotions, defined in a model called Hourglass of Emotions (see Figure 3).

3 Methodology

Typically, lexicon-based approaches for sentiment classification are based on the insight that the polarity of a piece of text can be obtained on the ground of the polarity of the words which compose it. However, due to the complexity of natural languages, a so simple approach is likely to fail since many facets of the language (e.g., the presence of the negation) are not taken into acccount. As a consequence, we propose a more fine-grained approach: given a Tweet T, we split it in several micro-phrases $m_1 \dots m_n$ according to the *splitting cues* occurring in the content. As *splitting cues* we used punctuations, adverbs and conjunctions. Whenever a *splitting cue* is found in the text, a new *micro-phrase* is built.

3.1 Description of the approach

Given such a representation, we define the sentiment S conveyed by a Tweet T as the sum of the polarity conveyed by each of the *micro-phrases* m_i which compose it. In turn, the polarity of each *micro-phrase* depends on the sentimental score of each term in the micro-phrase, labeled as $score(t_j)$, which is obtained from one of the above described lexical resources. In this preliminary formulation of the approach we did not take into account any *valence shifters* [7] except of the negation. When a *negation* is found in the text, the polarity of the whole micro-phrase is inverted. No heuristics have been adopted to deal with neither *language intensifiers* and *downtoners*, or to detect *irony* [14].

We defined four different implementations of such approach: BASIC, NORMALIZED, EMPHASIZED and EMPHASIZED-NORMALIZED. In the BASIC formulation, the

Fig. 3. The Hourglass of Emotions

sentiment of the Tweet is obtained by first summing the polarity of each micro-phrase. Then, the score is normalized through the length of the whole Tweet. In this case the micro-phrases are just exploited to invert the polarity when a negation is found in text.

$$S_{basic}(T) = \sum_{i=1}^{n} \frac{pol_{basic}(m_i)}{|T|} \tag{1}$$

$$pol_{basic}(m_i) = \sum_{j=1}^{k} score(t_j) \tag{2}$$

In the NORMALIZED formulation, the *micro-phrase-level* scores are normalized by using the length of the *single* micro-phrase, in order to weigh differently the micro-phrases according to their length.

$$S_{norm}(T) = \sum_{i=1}^{n} pol_{norm}(m_i) \tag{3}$$

$$pol_{norm}(m_i) = \sum_{j=1}^{k} \frac{score(t_j)}{|m_i|} \tag{4}$$

The EMPHASIZED version is an extension of the basic formulation which gives a bigger weight to the terms t_j belonging to specific POS categories:

$$S_{emph}(T) = \sum_{i=1}^{n} \frac{pol_{emph}(m_i)}{|T|} \tag{5}$$

$$pol_{emph}(m_i) = \sum_{j=1}^{k} score(t_j) * w_{pos(t_j)} \tag{6}$$

where $w_{pos(t_j)}$ is greater than 1 if $pos(t_j) = adverbs, verbs, adjectives$, otherwise 1.

Finally, the EMPHASIZED-NORMALIZED is just a combination of the second and third version of the approach:

$$S_{emphNorm}(T) = \sum_{i=1}^{n} pol_{emphNorm}(m_i) \tag{7}$$

$$pol_{emphNorm}(m_i) = \sum_{j=1}^{k} \frac{score(t_j) * w_{pos(t_j)}}{|m_i|} \tag{8}$$

3.2 Lexicon-based Score Determination

Regardless of the variant which is adopted, the effectiveness of the whole approach strictly depends on the way $score(t_j)$ is calculated. For each lexical resource, a different way to determine the sentiment score is adopted.

As regards *SentiWordNet*, t_j is processed through an NLP pipeline to get its POS-tag. Next, all the synsets mapped to that POS of the terms are extracted. Finally, $score(t_j)$ is calculated as the weighted average of all the *sentiment scores* of the sysnets.

If *WordNet-Affect* is chosen as lexical resource, the algorithm tries to map the term t_j to one of the nodes of the affective hierarchy. The hierarchy is climbed until a matching is obtained. In that case, the term inherits the sentiment score (extracted from SentiWordNet) of the A-Label it matches. Otherwise, it is ignored.

The determination of the score with *MPQA* and is quite straightforward, since the algorithm first associates the correct POS-tag to the term t_j, then looks for it in the lexicon. If found, the term is assigned with a different score according to its categorical label.

A similar approach is performed for *SenticNet*, since the knowledge-base is queried and the polarity associated to that term is obtained. However, given that SenticNet also models common sense concepts, the algorithm tries to match more complex expressions (as *bigrams* and *trigrams*) before looking for simple unigrams.

4 Experimental Evaluation

In the experimental session we evaluated the effectiveness of the above described lexical resources in the task of sentiment classification of microblog posts. Specifically, we evaluated the accurracy of our lexicon-based approach on varying both the four lexical resources as well as the four versions of the algorithm.

Dataset and Experimental Design: experiments were performed by exploiting SemEval-2013 [10] and Stanford Twitter Sentiment (STS) datasets [5]. SemEval-2013[1] dataset consists of 14,435 Tweets already split in training (8,180 Tweets) and test data (3,255). Tweets have been manually annotated and are classified as *positive*, *neutral* and *negative*. STS dataset contains more that 1,600,000 Tweets, already split in training and test test, but test set is considerably smaller than training (only 359 Tweets). In this case tweets have been collected through Twitter APIs[2] and automatically labeled according to the emoticons they contained.

Even if our approach can work in a totally unsupervised manner, we used training data to learn positive and negative classification thresholds through a simple Greedy strategy. For SemEval-2013 all the data were used to learn the thresholds, while for STS only 10,000 random tweets were exploited, due to computational issues. As regards the emphasis-based approach, the boosting factor w is set to 1.5 after a rough tuning (the score of adjectives, adverbs and nouns is increased by 50%). As regards the lexical resources, the last versions of MPQA, SentiWordNet and WordNet-Affect were downloaded, while SenticNet

[1] www.cs.york.ac.uk/semeval-2013/task2/
[2] https://dev.twitter.com/

was invoked through the available REST APIs[3]. Some statistics about the coverage of the lexical resources is provided is provided in Table 1. For POS-tagging of Tweets, we adopted TwitterNLP[4] [11], a resource specifically developed for POS-tagging of microblog posts. Finally, The effectiveness of the approaches was evaluated by calculating both accuracy and F1-measure [16] on test sets, while stastical significance was assessed through McNemar's test[5].

Lexicon	**SemEval-Test**	**STS-Test**
Vocabulary Size	*18,309*	*6,711*
SentiWordNet	4,314	883
WordNet-Affect	149	48
MPQA	897	224
SenticNet	1,497	326

Table 1. Statistics about coverage

Discussion of the Results: results of the experiments on SemEval-2013 data are provided in Figure 4. Due to space reasons, we only report *accuracy* scores. Results shows that the best-performing configuration is the one based on *SentiWordNet* which exploits both *emphasis* and *normalization*. By comparing all the variants, it emerges that the introduction of *emphasis* leads to an improvement in 7 out of 8 comparisons (0.4% on average). Differences are statistically significant only by considering the introduction of emphasis on normalized approach with SenticNet ($p < 0.0001$) and SentiWordNet ($p < 0.0008$). On the other side, the introduction of normalization leads to an improvement only in 1 out of 4 comparisons, by using the WordNet-Affect resource ($p < 0.04$). By comparing the effectiveness of the different lexical resources, it emerges that SentiWordNet performs significantly better than both SenticNet and WordNet-Affect ($p < 0.0001$). However, even if the gap with MPQA results quite large (0.7%, from 58.24 to 58.98), the difference is not statistically significant ($p < 0.5$). *To sum up, the analysis performed on SemEval-2013 showed that SentiWordNet and MPQA are the best-perfoming lexical resources on such data.*

Figure 5 shows the results of the approaches on STS dataset. Due to the small number of Tweets in the test set, results have a smaller statistical significance. In this case, the best-perfoming lexical resource is SenticNet, which obtained 74.65% of accuracy, greater than those obtained by the other lexical resources. However, the gap is statistically significant only if compared to WordNet-Affect ($p < 0.00001$) and almost significant with respect to MPQA ($p < 0.11$). Finally, even if the gap with SentiWordNet is around 2% (72.42% accuracy), the difference does not seem statistically significant ($p < 0.42$). Differently from SemEval-2013 data, it emerges that the introduction of *emphasis*

[3] http://sentic.net/api/
[4] http://www.ark.cs.cmu.edu/TweetNLP/
[5] http://en.wikipedia.org/wiki/McNemar's_test

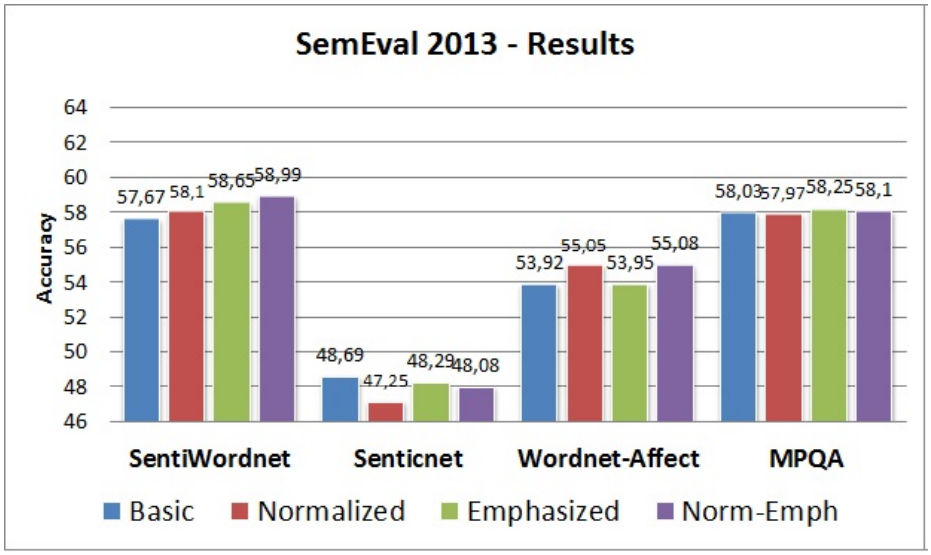

Fig. 4. Results - SemEval 2013 data

leads to an improvement only in 2 comparisons (+0.28% only on MPQA and WordNet-Affect), while in all the other cases no improvement was noted. The introduction of normalization produced a improvement in 3 out of 4 comparisons (average improvement of 0.6%, peak of 1.2% on MPQA). In all these cases, no statistical differences emerged on varying the approaches on the same lexical resource.

5 Conclusions and Future Work

In this paper we provided a thorough comparison of lexicon-based approaches for sentiment classification of microblog posts. Specifically, four widespread lexical resources and four different variants of our algorithm have been evaluated against two state of the art datasets.

Even if the results have been quite controversial, some interesting behavioral patterns were noted: **MPQA** and **SentiWordNet** emerged as the best-performing lexical resources on those data. This is an interesting outcome since even a resource with a smaller coverage as MPQA can produce results which are comparable to a general-purpose lexicon as SentiWordNet. This is probably due to the fact that subjective terms, which MPQA strongly rely on, play a key role for sentiment classification. On the other side, results obtained by **WordNet-Affect** were not good. This is partially due to the very small coverage of the lexicon, but it is likely that the choice of relying sentiment classification only on affective features filters out a lot of relevant terms. Finally, results obtained by **SenticNet** were really interesting since it was the best-performing configuration

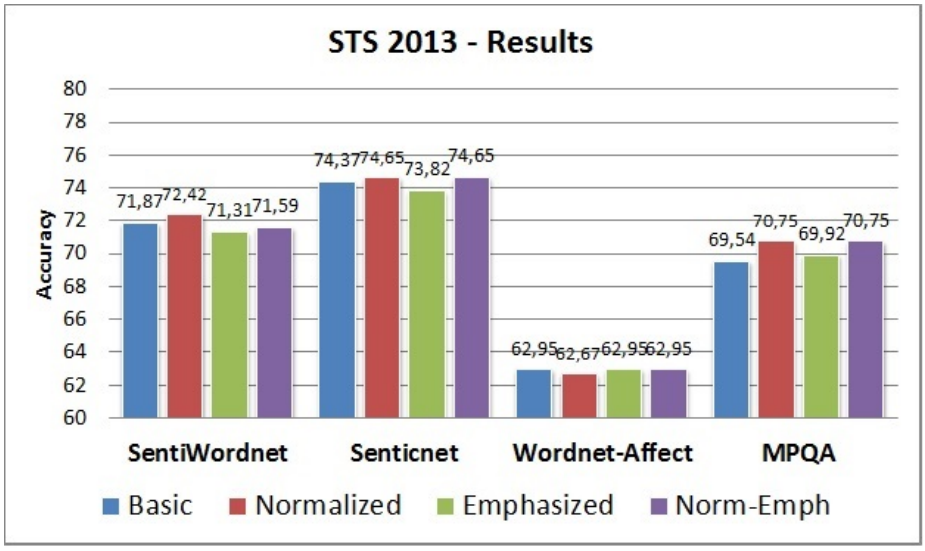

Fig. 5. Results - STS data

on STS and the worst-performing one on SemEval data. Further analysis on the results showed that this behaviour was due to the fact that SenticNet can hardly classificate neutral Tweets (only 20% accuracy on that data), and this negatively affected the overall results on a three-class classification task. Further analysis are needed to investigate this behavior.

As future work, we will extend the analysis by evaluating more lexical resources as well as more datasets. Moreover, we will refine our technique for threshold learning and we will try to improve our algorithm by modeling more complex syntactic structures as well as by introducing a word-sense disambiguation strategy to make our approach semantics-aware.

Acknowledgments. This work fullfils the research objectives of the project "VINCENTE - A Virtual collective INtelligenCe ENvironment to develop sustainable Technology Entrepreneurship ecosystems" funded by the Italian Ministry of University and Research (MIUR)

References

1. Andrea Esuli Baccianella, Stefano and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC*, volume 10, pages 2200–2204, 2010.
2. Erik Cambria and Amir Hussain. *Sentic computing*. Springer, 2012.
3. Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. *AAAI, Quebec City*, pages 1515–1521, 2014.

4. Xiaowen Ding, Bing Liu, and Philip S Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 231–240. ACM, 2008.
5. Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.
6. Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 537–546. ACM, 2013.
7. Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125, 2006.
8. Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, pages 415–463. Springer, 2012.
9. George A Miller. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
10. Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. 2013.
11. Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *HLT-NAACL*, pages 380–390, 2013.
12. Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, 2010.
13. Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.
14. Antonio Reyes, Paolo Rosso, and Tony Veale. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239–268, 2013.
15. Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. Semeval-2014 task 9: Sentiment analysis in twitter. *Proc. SemEval*, 2014.
16. Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.
17. Carlo Strapparava and Alessandro Valitutti. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086, 2004.
18. Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.
19. Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.