Opinion Learning without Emotional Words*

Marina Sokolova[1] and Guy Lapalme[2]

[1] Children's Hospital of Eastern Ontario Research Institute
msokolova@ehealthinformation.ca
[2] DIRO, Université de Montréal
lapalme@iro.umontreal.ca

Abstract. This paper shows that a detailed, although non-emotional, description of event or an action can be a reliable source for learning opinions. Empirical results show the practical utility of our approach and its competitiveness in comparison with previously used methods.

1 Motivation

Humans can infer opinion from details of event or action description: I saw it **several** times can signal a positive opinion in a movie review; there was a **long** wait may indicate a negative opinion about health care services or a positive opinion about entertainment. We show that, under certain conditions, quantitative (few, small) and stance (probably) indicators extracted from texts provide for successful machine learning. Opinion learning has mainly studied *polarity* of texts (I enjoyed this movie expresses positive polarity, I hated the film – negative polarity). With an increased supply of free-form, unstructured or loosely structured texts, learning opinions can benefit from an assessment of the parameters of the language, other than polarity.

We conjecture that descriptive words which emphasize quantitative properties (high, some), time (old, yesterday) and confidence in happening (can, necessary, probably) can be used in learning opinions. We organize descriptive words in a hierarchy: the lowest level works with words (the lexical level); the middle level generalizes word categories (semantics); the highest level applies to the text as a whole (pragmatics). The three levels are built upon quantitative and stance indicators found in text, e.g. stance/degree/time adverbs (probably,again), size/quantity/extent adjectives (long,many), degree pronouns (some). The hierarchy avoids the use of topic and emotionally-charged words. Whereas most sentiment analysis studies are stated as binary classification problems, we, instead, first assign texts numerical opinion tags (*regression*) and then classify them according to opinion categories (*classification*).

In previous work, we built a hierarchy on verbs which are not explicit opinion bearers, e.g., sit, eat [5]. We showed that the hierarchy worked well in environment where negative opinions were expressed indirectly. The empirical results have confirmed that such lexical features can provide reliable and, sometimes, better opinion classification than statistically selected features [5].

* Parts of this research were supported by NSERC funds available to both authors.

2 Hierarchical Text Representation

Previous studies of sentiment and opinion mining did not explicitly investigate non-contrived, often spontaneously written, text, e.g., forum posts. Based on [2], we suggest that some lexical characteristics of non-contrived text are worth exploring : (i) a larger proportion of descriptive degree pronouns (some, few, none); (ii) frequently used adverbs of time (today), degree (roughly), stance (maybe), frequency (often); (iii) a larger proportion of descriptive adjectives, e.g., size/quantity/extent (big), time (new), relational (same); their gradable derivatives (biggest); (iv) frequent use of order words, e.g. ordinal and cardinal numbers (first, one); (v) frequent use of stance verbs, i.e. modal verbs (can) and mental verbs (think).

We use those word categories to build the lower level for the hierarchical text representation. We find seed words in [2] and add their synonyms from an electronic version of Roget's Interactive Thesaurus[3]. To accommodate negative comments, we added negations. We ignore derived, topical, affiliative, foreign words which are less frequent in non-contrived text. We purposefully omit evaluative/emotive adjectives. This omission allows emphasis on the role of quantitative description in text.

Starting from the bottom, the hierarchy defines the word categories used in detailed descriptions, then groups the categories into four types of comments, and finally combines the types into direct and indirect detailed categories:

directDetails presents primary clues of quantitative evaluation and attributes of the discussed issues:

- *Estimation* lists the reference attributes: physical parameters, relative and absolute time (corresponding adverbs and adjectives).
- *Quantification* expresses the broadness of the discussed reference (adverbs, adjectives, pronouns of degree, cardinal numbers, frequency adverbs);

indirectDetails presents secondary clues of the issue evaluation:

- *Comparison* presents a comparative evaluation of the discussed issues, their qualities and relations among them (gradable relation adjectives and ordinal numbers).
- *Confidence* reflects on the certainty about the happening of events (stance adverbs and verbs from the lower level).

Our assumption is the following: descriptive words emphasize important characteristics of the discussed issues. In sentences, such words usually precede their references, especially in non-contrived text [2]. Hence, the extraction of words which follow the descriptors results in the set of words most emphasized in the text. The idea behind the extraction procedure is the following: two-word sequences - bigrams - which have descriptors on their left side capture the modified and intensified words. After extracting such bigrams, we find modified and intensified words. The probability of the word occurrence after a descriptor reveals frequently modified and intensified words. Concentrating on one-side bigrams prevents the multiple extraction of the same word.

[3] http://thesaurus.reference.com/

Table 1. SVM's results in learning of opinions. The features sets are defined in Section 3. The best results are in **bold**. For regression, the mean overall value baseline: *Relative Absolute Error* (*rae*) – 100.2, *Root Relative Squared Error*(*rrse*) – 101.1. For classification, the majority class baseline: *Accuracy*(*Acc*) – 52.6%

	Regression						Classification (%)							
	Text Features						Text Features							
	I		**II**		**III**		**I**		**II**		**III**			
											numeric		binary	
Opinion	*rae*	*rrse*	*rae*	*rrse*	*rae*	*rrse*	*Acc*	*R*	*Acc*	*R*	*Acc*	*R*	*Acc*	*R*
summed	95.1	97.6	91.4	94.4	82.0	86.7	67.3	72.9	68.4	75.8	70.1	82.4	73.6	78.2
positive	75.8	82.6	77.7	84.4	**63.7**	70.6	73.3	79.4	75.0	81.2	76.1	83.7	**79.0**	80.3
negative	89.4	88.1	87.2	85.8	75.17	80.2	63.7	67.3	66.6	69.7	67.8	73.9	71.3	76.6

3 Empirical Results

We ran experiments on consumer-written reviews which exemplify non-contrived text. 314 free-form commentaries which segments are marked with a mixture of positive and negative opinion tags were introduced in [3]. The extensively labeled texts are a more realistic, albeit difficult, environment for machine learning than those which are tagged with positive/negative tags.

For regression problems, for each text, we computed three numerical labels: the number of positive tags; the number of negative tags; a signed sum of the two numbers. To state classification problems, we apply unsupervised equal-frequency discretization to each of the three label sets [4]. The resulting classification problems could be dubbed as *stronger* vs *weaker* classes.

We construct three feature sets for text representation: **I**, direct descriptors enhanced by the most frequent extracted words; cut-off was determined by frequencies of personal pronouns; **II**, all descriptors enhanced by the most frequent extracted words; **III**, all the extracted words with frequency > 5, with numerical and binary attribute versions for the classification problems. Attribute normalization with respect to the number of words in the text eliminates the length bias. We applied Support Vector Machine, K-Nearest Neighbor and decision-based M5P Trees (regression) and REP Trees (classification) and their bagged versions. [4] For all the learning problems, SVM was considerably more accurate than the other algorithms. Table 1 reports SVM's best results (ten-fold cross-validation).

In regression problems, for both measures, the best results reported in this work are *statistically significant* better the previous best results (*rae*: paired t-test, $P = 0.0408$; *rrse*: paired t-test, $P = 0.0418$)[5]. In classification problems, current best results for *Recall* provide *very statistically significant* improvement (paired t-test, P=0.0071), whereas *Accuracy* improvement is not statistically significant (paired t-test, P=0.6292).

[4] http://www.cs.waikato.ac.nz/ml/weka/

4 Discussion and future work

Machine learning algorithms were successfully used in opinion classification [6]. Some of this work relied on characteristics of reviewed products, e.g. main product characteristics were extracted based on association rule mining algorithms in conjunction with feature frequency [3]. In contrast, we opted for a method which does not involve the use of the domain's content words. Syntactic and semantic features that express the intensity of terms were used to classify opinion intensity [7]. We, instead, focus on the formalization and utilization of non-emotional lexical features. Whereas [1] compared the use of adverbs and adjectives with adjectives only, we concentrated on descriptive adjectives and adverbs.

In our experiments, we studied the relevance of detailed, specific comments to the learning of positive, negative, and summed opinions in both *regression* and *classification* settings. Learning results of positive opinion turned out to be more accurate. Hence relations between features representing descriptive details and positive opinion are easier to detect than those for other problems. The obtained empirical results show improvement over baselines and previous research. The improvement especially significant for regression problems.

Our approach can be applied to analyze language of opinions for other types of texts, for example, blogs and reviews published in traditional media. The first step would be to determine the distinctive word categories of such texts, and then find the semantic parameters of opinion disclosure which use them. Next, we can use the formalized patterns for information extraction. On the last phase, we would use the extracted information to represent texts in machine learning experiments. The application of our method could result in finding similarities imposed by shared communication characteristics, e.g., the one-to-many speaker-audience interaction, and differences that can be attributed to dissimilar characteristics, e.g., Internet and published media.

References

1. Benamara, F., C. Cesarano, A. Picariello, D. Reforgiato, and V. Subrahmanian. 2007. Sentiment analysis: Adjectives and adverbs are better than the adjectives alone. *Proceedings of ICWSM'2007.*
2. Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *Longman Grammar of Spoken and Written English.* Longman.
3. Hu, M. and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the KDD'04*, pages 168–177.
4. Reinratz, T. 1999. *Focusing Solutions for Data Mining.* Springer.
5. Sokolova, M. and G. Lapalme. 2008. Verbs Speak Loud: Verb Categories in Learning Polarity and Strength of Opinions. In *Proceedings of Canadian AI'2008*, pages 320–331. Springer.
6. Sokolova, M. and S. Szpakowicz. 2009. Machine Learning Application in Mega-Text Processing. E. Soria, J. Martin, R. Magdalena, M.Martinez, and A. Serrano (ed.) *Handbook of Research on Machine Learning Applications*, IGI Global.
7. Wilson, T., J. Wiebe, and R. Hwa. 2006. Recognizing strong and weak opinion clauses. *Computational Intelligence*, 22(2):73–99.