Considering Impact of Sociolinguistic Findings in Believable Opinion Mining Systems

Alexander Osherenko, Socioware Development, osherenko@socioware.de

1 Introduction

Opinions are frequent means of communication in human society and automatic approaches to opinion mining in texts attracted therefore much attention. All in all, most approaches apply data mining techniques and extract lexical features (words) as reliable means of classification [1].

Noteworthy that although the interest in opinion mining is huge, there are only few explorations on words extracted in opinion mining. This study considers this drawback and elaborates on a sociolinguistic explanation. We hypothesize: an opinion mining system should be trained for classifying opinions in texts of the same language style[2].

Hence, this contribution focuses on the following questions: 1) do sociolinguistic aspects of corpora, for example, their colloquiality or literariness, influence classification results; 2) how should reliable opinion mining systems train to obtain trustworthy classification results.

2 Corpora

In our study, we use 4 text corpora of the same (emotional) domain: the Sensitive Artificial Listener (SAL) corpus [4], the Berardinelli movie reviews corpus (BMRC), Pang movie reviews corpus (PMRC) [5], the Corpus with product reviews (CwPR). The SAL corpus consists of 27 natural-language dialogues and includes short utterances (less than 200 words) with repetitions, repairs, vague formulation annotated with 5 emotional classes. The own BMRC corpus comprises of 215 movie reviews [2] annotated with 9 emotional classes that are grammatically correct, long texts. The PMRC corpus containing 2000 weblogs annotated with 2 emotional classes are movie reviews that can be short texts with vague formulation. The own CwPR corpus that contains 300 product reviews fetched from the Internet [3] is annotated using 5 emotional classes. Hence, the "most" literary corpus is BMRC, followed by in descending order of literariness CwPR, PMRC and the "least" literary SAL; the "most" colloquial corpus is SAL, followed in descending order of colloquiality by PMRC, CwPR and the "least" colloquial BMRC.

[2]Adopted to texts, Belikov and Krisin [6] define language style in sociolinguistics roughly as texts that can be distinguished, for example, by language code, natural (English) or artificial (Esperanto) or by language subcode, literary or colloquial language.

3 Results

We compose a non-sociolinguistic dataset and a sociolinguistic dataset each for all corpora (overall 7 datasets). The non-sociolingustic datasets are datasets with all lexical features from the frequency list of a corpus, for example, the non-sociolingustic PMRC dataset contains the whole frequency list of 38476 lexical features. In the sociolinguistic datasets, we: 1) limit number of extracted lexical features in datasets from the list of most frequent words to equal number 2038 from the SAL corpus to facilitate classification and avoid overfitting, for example, in sociolinguistic PMRC dataset we omit 36438 features from the non-sociolinguistic dataset; 2) use the classes-number recognition value to measure classification success and to normalize results in context of different number of classes in the corpora.

Success of classification in data mining is measured typically using conventional recall or precision values. However, in the current case these measures do not provide explicit consideration of number of annotated classes and therefore no trustworthy basis to facilitate results' comparison of different corpora.

Corpus	$R_0(\%)$	$CN_0(\%)$	$R(\%)$	$CN(\%)$
SAL	60.2	83.5	60.2	83.5
PMRC	86	83.7	81.6	77.4
CwPR	51	76	48.7	73.6
BMRC	31.9	73.3	28	67.9

Table 1: Results

Thus, as an appropriate measure we use the classes-number recognition value (see [1] for exact definition):

$$CN(R) = \frac{NR - 1}{(N-1)R} \quad (1)$$

where R is the recall value averaged over classes, $N > 1$ is number of classes in a corpus. Note that $CN(R) \in [0..1]$.

We classify corpora with the SVM classifier from the WEKA toolkit using 10-fold cross-validation [7] and evaluate lexical features using the presence feature method. We choose the SVM classifier as an analytical classification method to highlight sociolinguistic differences between recognition results (Table 1).

Table 1 shows results of classification of non-sociolinguistic and sociolinguistic datasets of different corpora (column *Corpus*) where R_0 and R values — recall values averaged over classes for the non-sociolinguistic and sociolinguistic datasets respectively; the CN_0 and CN columns specify the corresponding classes-number values.

4 Discussion & Future Work

Consideration of sociolinguistic aspects influences classification results. First, the order of classification values of non-sociolinguistic datasets defined by the CN_0 value (PMRC, SAL, CwPR, BMRC) swaps SAL and PMRC and results after sociolinguistic consideration in order (SAL, PMRC, CwPR, BMRC) sorted by the CN value.

Second, sociolinguistic consideration allows comprehensive interpretation of the CN value (in contrast to non-sociolinguistic CN_0). Hence, the "most" colloquial corpus (SAL) calculates the maximal value CN value, 83.5%; the "most literary" corpus (BMRC) — the minimal value, 67.9% whereas PMRC is "less" colloquial than SAL and "less" literary than CwPR (77.4%) and CwPR is "more" literary than PMRC and "more" colloquial than BMRC (73.6%). Note that conventional R values do not allow such interpretation.

Hence, we assume that reliable software systems performing opinion mining in texts should consider sociolinguistic aspects of feature extraction. Our naive explanation: consideration, for example, of the same language style would increase possibility that two independent feature sets contain the same features that are properly initialized after training.

We assume that sociolinguistic aspects influence classification in systems relying on other modalities, for example, acoustic — the system that recognizes opinions in official speech would fail to reliably classify opinions in conversations between friends. Thus, we will explore other corpora to verify our assumption. Moreover, we will elaborate on composing universal set of features that can be extracted for classification of any corpus.

References

[1] А.М. Ошеренко. *Opinion mining and lexical affect sensing. Computer-aided analysis of opinion and emotions in texts.* PhD thesis, Südwestdeutscher Verlag für Hochschulschriften, 2011.

[2] J. Berardinelli. Reelviews movie reviews, 2010. http://www.reelviews.net.

[3] Epinions. Product reviews, 2008. http://www.epinions.com.

[4] S. Kollias. Ermis project, 2007. http://www.image.ntua.gr/ermis/.

[5] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

[6] В.И. Беликов, Л.П. Крысин. *Социолингвистика.* Изд. РГГУ, 2001.

[7] I. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, second edition, 2005.