Is Sentiment a Property of Synsets? Evaluating Resources for Sentiment Classification using Machine Learning

Aleksander Wawer

Institute of Computer Science, Polish Academy of Science.
ul. J.K. Ordona 21, 01-237 Warszawa. Poland
axw@ipipan.waw.pl

Abstract

Existing approaches to classifying documents by sentiment include machine learning with features created from n-grams and part of speech. This paper explores a different approach and examines performance of one selected machine learning algorithm, Support Vector Machines, with features computed using existing lexical resources. Special attention has been paid to fine tuning of the algorithm regarding number of features. The immediate purpose of this experiment is to evaluate lexical and sentiment resources in document-level sentiment classification task. Results described in the paper are also useful to indicate how lexicon design, different language dimensions and semantic categories contribute to document-level sentiment recognition. In a less direct way (through the examination of evaluated resources), the experiment analyzes adequacy of lexemes, word senses and synsets as different possible layers for ascribing sentiment, or as candidates for sentiment carriers. The proposed approach of machine learning word category frequencies instead of n-grams and part of speech features can potentially exhibit improvements in domain independency, but this hypothesis has to be verified in future works.

1. Background

In the light of recent rise of interest in quantifying sentiment in language, it is increasingly more important to understand how various approaches to lexicon and ontology design contribute to relative success or failure of automated sentiment recognition. Exploring lexical rather than syntactic aspects of sentiment analysis seems justified for at least two reasons. Firstly, successes of bag of words or unigram based methods (**?**) indicate predominant role of the lexical dimension in sentiment analysis. Secondly, the recognition of syntax structures identified as important for sentiment polarization (Agarwal et al., 2008) does not contribute much to document-level classification accuracy, especially when applied to large document collections.

Since decades it is known that words can be grouped along semantic axes that cover a corresponding dimension (Deese, 1964). General Inquirer (GI) and Dictionary of Affect (DAL) lexicons were based on that notion. Another way of grouping words is according to their equivalence for the purposes of information retrieval: this idea defines WordNet's synsets. Both ways of grouping express certain type of similarity between words within the same group. By computing frequencies of lexical units belonging to each group, one can attempt to classify document level sentiment using machine learning on frequency distributions of these groups. Experiments presented in this article are aimed at investigating usefulness of those groupings for sentiment classification, and thus evaluating associated language resources. Our way is then prior-knowledge-free in terms of unknown distributional properties of the groups between document classes and also knowledge-based, because we use existing resources to compute group membership frequencies.

The work presented in this article attempts to address several problems:

- Investigate how various word meta-categories and groupings perform when applied to sentiment analysis. What dimensions (semantic axes) of language are the most relevant for this task? How are they related to document level sentiment?
- Review and evaluate existing resources and sentiment lexicons using a common benchmark.

2. Evaluation

To evaluate the performance of lexical resources under homogenous criteria, we have focused on one machine learning algorithm of choice, namely Support Vector Machines (SVM). Firstly, its behaviour has been thoroughly studied for bag of words document classification problems (Colas et al., 2007), with special attention paid to feature vectors size and training sets size. As it will be shown below, feature vector sizes used in the comparison differ by several orders of magnitude, which demans tuning of the alogrithm to perform optimally. Secondly, in many reported text classification experiments including seminal work of (Pang et al., 2002) SVM outperformed comparable techniques. The experiments presented below were conducted on well-known Sentiment Polarity Dataset 2.0[1]. For each feature set (and thus each resource) described below, we compared average classification accuracy obtained in three-fold cross validation.

3. Support Vector Machines

Support Vector Machines (SVM) method has at least one parameter proven crucial (Colas et al., 2007) for experiments with different feature vector sizes, namely the free parameter C. In the SVM method, non-linear separability problems are solved by introducing ξ variables to:

$$\frac{\|w\|^2}{2} + C\sum_{i=1}^{n}\xi_i$$

The dual form is obtained by posing a Lagrangian:

$$W(\alpha) = \sum\nolimits_{i=1}^{N}\alpha_i - \frac{1}{2}\sum\nolimits_{j=1}^{N} y_i y_j \alpha_i \alpha_j \left(x_i \cdot x_j\right)$$

[1] http://www.cs.cornell.edu/People/pabo/movie-review-data/

subject to constraints:

$$\sum\nolimits_{i=1}^{N} y_i \alpha_i = 0$$

$$0 \leq \alpha_i \leq C, i = 1...N$$

Document-level sentiment classification experiments using SVM (Colas et al., 2007) indicate a relationship between the C parameter, feature size vector and classification accurracy. Because of that, evaluation of lexical resources with different feature vector sizes has to take into account a range of C values[2].

4. Lexical Resources

Compared lexicons differ as to their origins, age and most notably, design. Interestingly all of the resources, perhaps except SentiWordNet, were designed by psychologists. Attitudes and emotions are attributable to word senses (GI), sets of senses (SentiWordNet) or lexemes (DAL). The evaluation of lexicons presented in this paper is not fully comprehensive, but probably takes into account most of the established and recognized resources. Specifically, some of the lexicons described in the literature, such as the Clairvoyance affect lexicon (Grefenstette et al., 2006), were not considered due to their limited availability.

4.1. General Inquirer (GI)

General Inquirer [3] default dictionary, which is consists of Harvard IV and Laswell dictionaries, contains 11767 word senses. Most of the dictionary words have one sense and granularity of meanings is lower than in WordNet. Each sense is mapped to over 180 categories, which include several emotion-related ones such as pleasure, pain, feelings or arousal. Probably the two most relevant categories are Positive and Negative, Osgood's evaluative dimension. Category membership is binary: word sense either belongs to a category or not.

For text processing, GI application uses a dictionary-backed lemmatizer and word sense disambiguation using Kelly/Stone rules (Kelly and Stone, 1975). Thus, on the General Inquirer output, sentiment (appropriate word tags) is a property of lexeme senses. Currently, General Inquirer's lexicon is often used as Gold Standard in various attempts of automated acquisition of sentiment related vocabulary (**?**).

4.2. Dictionary of Affect in Language (DAL)

Cynthia Whissel's Dictionary of Affect in Language (DAL) (Whissel, 1989) is a resource newer than GI, developed manually to measure emotional meanings of words and texts. It contains 8742 words scored on three categories: evaluation (also called pleasantness), activation and imagery on a continuous scale ranging between 1 and 3. For some reason, DAL never received as much attention from the sentiment analysis community as the General Inquirer. Notable exceptions include the recent work of (Agarwal et al., 2009).

4.3. WordNet (WN)

WordNet is one of the most well-known and established resources in language processing (Fellbaum, 1998; Miller et al., 1990); As of 2006, the database contained about 150,000 words organized in over 115,000 synsets for a total of 207,000 word-sense pairs. WordNet does not contain explicit sentiment information and it is not feasible to use it directly for computing sentiment scores.

However, WordNet synsets can be used as the most fine grained way of grouping words. The intuition is that the same sense can appear in different reviews (documents) and possibly even expressed with different words — nevertheless denoting a single concept. If sentiment is attributable to synsets, as it is assumed in (Esuli and Sebastiani, 2006), then focusing on synset occurrences in text should improve sentiment recognition, obviously assuming that synset-level sentiment is known. And this can be so, if the synset appeared in training part of the dataset and its sentiment is reflected in trained classifiers. However, using WordNet synset occurrences in unrestricted way by taking all possible synsets without disambiguation, leads to very large number of features. In our approach, as will be detailed below, we mapped each lexeme (lemmatized unigram) to a set of possible synsets, first without any disambiguation, learning all excessive senses, then restricting the senses to those that agree with POS, as detected for a given word.

4.4. SentiWordNet (SWN)

SentiWordNet (Esuli and Sebastiani, 2006) is the newest resource among the tested. It has been built on top of WordNet 2.0 with three key assumptions:

- sentiment is attributable to synsets,
- synsets can exhibit both positive and negative properties simultaneously,
- presence of positive or negative affect makes synsets subjective.

The method used to assign sentiment scores to WordNet synsets was semi-supervised classification of glosses associated to synsets. In the machine learning experiments described below we applied SentiWordNet version 1.0, the most recent available. Because SentiWordNet contains synset sentiment scores, computing document-level score is as simple as an aggregation of sentiment over synsets identified in a document.

5. Features

This section contains descriptions of feature sets used in the experiment.

- **GI/PN**: only Positive and Negative categories from the General Inquirer, Kelly/Stone disambiguation applied.
- **GI/ALL**: summed occurrences of each General Inquirer dictionary (Harvard IV and Laswell) category, Kelly/Stone disambiguation applied.
- **DAL**: summed occurrences of each DAL dictionary category.

[2] The set of C values used in this paper follows (Colas et al., 2007).

[3] `http://www.wjh.harvard.edu/~inquirer/`

- **WN**: summed occurrences of each possible WN synset.
- **WN/POS**: summed occurrences of each WN synset, restricted to synsets matching part of speech, as recognized in text.
- **WN/POS2**: summed occurrences of each WN synset, restricted to synsets matching part of speech if recognized in text, all possible synsets otherwise.
- **SWN**: average positive and negative load of all SWN synsets, restricted to synsets matching POS, as recognized in text.
- **UG**: (binary) unigram occurrences, as in (Pang et al., 2002).

Additionally, low-dimensional feature sets (GI/PN, GI/ALL, DAL, SWN) have been extended with one more special feature, namely length of a document as number of tokens. The introduction of this feature can enable differentiation of classifications (in this case, support vectors) within movie reviews of different sizes.

6. Results

The results were obtained using Joachim's SVMLight[4] implementation of Support Vector Machines algorithm. In order to facilitate comparisons with (Pang et al., 2002), all parameters except C (where the best value has been selected, as in (Colas et al., 2007)) were left out at their default settings. The table 1 presents average accuracy in 3-fold cross-validation and number of features.

7. Discussion

We have confirmed the relationship between accuracy, feature vector size and SVM free parameter C, as reported by (Colas et al., 2007). Results obtained for SWN, DAL and GI/PN do not support the statement, expressed in (Colas et al., 2007), that C has a negligible effect in low dimensional feature sets. The default C setting in SVMLight, computed as $[avg.x * x]^-1$, typically yields suboptimal results. Only in the case of UG (unigram) features, this setting was found to be optimal. The differences obtained using different C indicate that the values reported in the foundational experiment of (Pang et al., 2002) could be probably improved upon.

It appears that GI/ALL feature set is a reasonable compromise between the number of features (180) and classification accuracy (78%). Two General Inquirer categories alone, Positiv and Negativ (GI/PN), achieved nearly 66% accuracy. This was the highest performing low dimensional feature set. What is interesting is the slight supremacy of DAL over SWN, especially given that the closest to evaluative dimension of DAL (which is Pleasantness) is not as explicit in encoding sentiment as SentiWordNet's Positive and Negative scores. It is not possible to explain comprehensively what caused these differences without further in-depth studies. Perhaps DAL's superiority was caused by its manual origin, as compared to SentiWordNet's semi-supervised, largely automated development.

Alternatively, the problem could be more fundamental: the relation of contextual truth-preserving interchangeability, which defines what synsets are, may not be a proper carrier for connotative equivalence. Thus, sentiment may not be attributable to synsets. Terms or words belonging to the same synset may have different or even opposing sentimental connotations, while remaining in the relation of semantic equivalence (belonging to the same synset). This point can be illustrated by two variants of one well-known subjective sentence, (Wiebe and Mihalcea, 2006):

- *That doctor is a quack.*
- *That doctor is medically unqualified.*

After replacing one word with its synonym, the sentence conveys quite different connotations.

Globally, the best performing feature set (WN) was also the richest in terms of the amount of information, which included a great deal of noise and redundancy. WordNet based feature set disambiguated with POS (WN/POS) achieved 5% lower accuracy. This indicates that the machine learning method applied in our experiments benefits from noise and reduntant information rather than suffers from the incorrectness thus introduced.

8. Further Work

Several problems drawn in this study need to be answered. The first one is related to appropriate sentiment binding (word, synset or perhaps word sense level) and related optimal design of sentiment language resources. The other problem that needs subsequent studies is whether and how sentiment analysis, especially document level sentiment classification using machine learning approaches, can benefit from word sense disambiguation. The issue that requires addressing is the fact that reduction in the number of features has a stronger decreasing effect on classification precision than improvement introduced by removing redundant word senses.

[4] `http://svmlight.joachims.org`

C	GI/PN	GI/ALL	DAL	UG	WN	WN/POS	WN/POS2	SWN
0.2	47.93	66.92	52.2	85.06	87.21	81.30	81.44	58.68
0.1	50.70	75.64	57.65	85.06	87.21	81.30	81.24	**58.68**
0.05	58.98	76.49	60.04	85.06	87.21	81.36	81.38	58.47
0.01	**65.68**	76.63	60.04	85.00	87.06	**82.06**	**82.74**	58.25
0.005	65.63	77.64	59.97	**85.67**	**87.31**	81.55	82.73	57.89
0.001	65.73	77.88	**60.19**	83.49	86.70	79.85	81.97	54.03
0.0005	65.63	**78.14**	59.69	76.32	85.44	78.35	80.99	53.96
0.0001	64.91	76.94	59.27	48.71	80.18	73.42	76.33	51.78
0.00001	61.58	70.32	57.6	48.71	65.26	53.91	64.96	48.94
0.000001	56.58	58.96	56.38	48.71	49.35	49.60	50.01	49.01
default	57.69	61.89	55.78	**85.74**	80.13	78.13	76.98	55.12
features	3	182	4	49898	103132	30567	35705	3

Table 1: Average three-fold cross-validation SVM accuracies for selected values of C, in percent. Best C setting for a given feature set (column) marked with bold fonts.

9. References

Ritesh Agarwal, T. V. Prabhakar, and Sugato Chakrabarty. 2008. I know what you feel: Analyzing the role of conjunctions in automatic sentiment analysis. *Advances in Natural Language Processing (Springer LNCS 5221/2008).*

Apoorv Agarwal, Fadi Biadsy, and Kathleen McKeown. 2009. Dal: Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceeding of EACL 2009. Athens, Greece.*

Fabrice Colas, Pavel Paclik, Joost N. Kok, and Pavel Brazdil. 2007. Does svm really scale up to large bag of words feature spaces. *Advances in Intelligent Data Analysis VII, Springer LNCS 4723/2007.*

James Deese. 1964. The associative structure of some common english adjectives. *Journal of Verbal Learning and Verbal Behavior*, 3(5).

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC.*

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database.* MIT Press.

Gregory Grefenstette, Yan Qu, David A. Evans, and James G. Shanahan, 2006. *Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes.* Springer. Netherlands.

Edward Kelly and Philip Stone. 1975. *Computer Recognition of English Word Senses.* North-Holland, Amsterdam.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: an on-line lexical database*. *International Journal of Lexicography*, 3(4):235–244, January.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP.*

Cynthia Whissel, 1989. *The dictionary of affect in language*, volume 4. Academic Press, London.

Janyce Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1065–1072.