Modelling SO-CAL in an Inheritance-based Sentiment Analysis Framework

F. Sharmila Satthar

University of Brighton
Watts Building, Lewes Road, Brighton, UK
F.Satthar@brighton.ac.uk

Abstract

Sentiment analysis is the computational study of people's opinions, as expressed in text. This is an active area of research in Natural Language Processing with many applications in social media. There are two main approaches to sentiment analysis: machine learning and lexicon-based. The machine learning approach uses statistical modelling techniques, whereas the lexicon-based approach uses 'sentiment lexicons' containing explicit sentiment values for individual words to calculate sentiment scores for documents. In this paper we present a novel method for modelling lexicon-based sentiment analysis using a lexical inheritance network. Further, we present a case study of applying inheritance-based modelling to an existing sentiment analysis system as proof of concept, before developing the ideas further in future work.

1998 ACM Subject Classification I.2.7 Natural Language Processing

Keywords and phrases Sentiment analysis, NLP, Inheritance network, Lexicon-based

Digital Object Identifier 10.4230/OASIcs.ICCSW.2015.46

1 Introduction

Sentiment analysis is the computational study of people's opinions, as expressed in text. It is important when a company or service provider wants to understand their users' needs, or share users' opinions and reviews about products or services with other potential users [8] [6]. A commonly used way of detecting sentiment is to caluclate 'semantic orientation' (SO), a numeric measure of subjectivity and opinion in text, for example in film reviews [13]. In machine learning approaches to sentiment analysis, semantic orientation scores are learned using statistical modelling from prepared training data. In lexicon-based approaches, semantic orientation scores are associated with individual words (such as +3 for 'good', −3 for 'bad'), and the total score for a text is calculated using heuristic rules. Key advantages of the lexicon-based approaches is that they do not require preparation of extensive training data sets, and their heuristic rules can utilize linguistic context to determine the sentiment of complex constructions, for example valence shifters [9], [5] such as intensifiers or negators. The major challenge for lexicon-based methods is coverage – handling words that are not in the lexicon or constructions that were not predicted by the rule designers.

One branch of previous research in natural language lexicons makes use of non-monotonic (default) inheritance networks to represent lexical information [1]. Regular and irregular words can be represented in a hierarchical structure with abstraction that shares common properties and behaviours, but also allows irregular words to specify only those aspects that deviate from the regular case. Our intuition is that lexicon-based sentiment analysis systems can be made more accurate by using such default inheritance-based lexical knowledge representation, and that this approach will allow us to address some of the coverage limitations of previous approaches. As a first step towards this goal, in this paper we present our work on modelling

2015 Imperial College Computing Student Workshop (ICCSW 2015).
Editors: Claudia Schulz and Daniel Liew; pp. 46–53

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

an existing lexicon-based approach to sentiment analysis in an inheritance-based framework using the lexical representation language DATR [4].

The system we modelled was Taboada el al.'s 'semantic orientation calculator' (SO-CAL) [11] [12] [10]. In SO-CAL, sentiment is represented by the semantic orientation of the text, which is expressed as both the word's (semantic) polarity and its strength (intensity). So, a semantic orientation score of a word/text determines whether it is positive or negative depending on its sign and how strong it is depending on its magnitude. SO-CAL uses a pure lexical method in which they calculate semantic orientation of a text by aggregating the semantic orientation of each opinion word present in the text, applying various heuristic rules to take account of contextual constructions.

In section 2, we briefly describe SO-CAL and the features used in its heuristics. In section 3, we describe the inheritance-based framework and our sentiment analysis system, Galadriel. In section 4, we describe how we model SO-CAL in Galadriel and in section 5, we present an evaluation that shows how Galadriel's performance compares with SO-CAL. Finally, section 6 contains discussion and future work.

2 The SO-CAL system

In SO-CAL [11], Taboada et al. aimed to analyse semantic orientation of individual words and contextual valence shifters in depth. However, they did not focus on linguistic analysis. First they extracted sentiment-bearing words including adjectives, adverbs, nouns and verb in a document. Then they used the semantic orientation score for each word from semantic orientation dictionaries to calculate a score for the whole document, taking into account valence shifters such as intensifiers and negators. Semantic orientation dictionaries are special dictionaries which include words with their semantic orientation. Taboada et al. created their dictionaries manually, as they believed that the way of creating dictionaries affects the overall accuracy of final results. Therefore, as a first step they created dictionaries for the words, which contain adjectives, adverbs, verbs and nouns with sentiment scores between $+5$ and -5 ($+$ sign refers to the positive polarity and $-$ sign refers to the negative polarity, and a semantically neutral word has a zero score).

To obtain the semantic score for a given document, SO-CAL calculates the overall SO value by adding together the semantic scores of words present in the document. In addition, for various classes of words, rules are invoked to modify the SO scores. For example, intensifier words modify the SO score of the word they are attached to (eg 'very good'): the SO dictionary specifies how big this modification is for each intensifier as a percentage (so 'very' involves a bigger change than 'slightly'). As well as intensifiers, SO-CAL has rules for negators, irrealis (hypothetical statements), repetition and positive bias which are explained in more detail below. This approach makes two key assumptions: the semantic orientation of a word is independent of its (broader) context and the semantic orientation can be expressed in numerical value.

3 The Inheritance-based Framework

In order to model SO-CAL, we start out with the 'Galadriel', which is a sentiment framework using the DATR/ELF lexicon representation system. DATR/ELF is a default inheritance based language processing system. DATR/ELF can encode very complex lexical information relating to phonology, morphology, syntax and semantics. Our research aims to exploit this information for sentiment analysis.

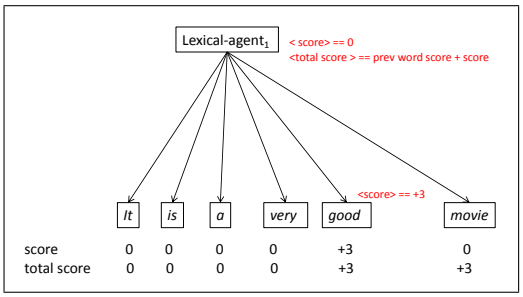

Figure 1 Simple sentiment model: add up raw sentiment score of all words.

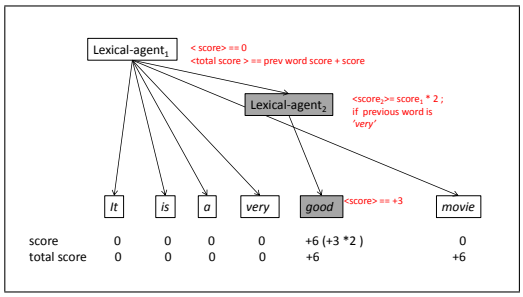

Figure 2 Sentiment model with intensifiers: *'very'* changes sentiment score of following word.

3.1 DATR and ELF

Evans and Gazdar [4] designed a lexical description language, DATR, to model the structure of the lexicon using default inheritance to capture complex class, subclass and exception relationships between words. More recently, Evans [3] introduced the Extended Lexicon Framework (ELF), a development which uses DATR to represent words not as isolated individuals, but as instances occurring in sentences. In ELF, information is still represented on a word-by-word basis, but the information about a word can depend upon information about its neighbours in a sentence. This allows ELF to represent more complex properties of whole sentences, while retaining its default lexical character to express exceptional cases.

ELF is based on two core ideas: the first is to view each word as a 'lexical agent', containing fixed information about the word itself, represented as features with values, and rules for calculating more complex values. These rules can refer to other features of the word, but also to features of adjacent words when the word appears in a sentence. For example the lexical agent for 'a' can look at the word to its right to decide whether its form feature should be 'a' or 'an'. The second is that these specifications of values and rules for lexical agents are organised into a default inheritance hierarchy, so that words with similar behaviour share the rules defining that behaviour.

3.2 The Basic Galadriel System

The basic Galadriel system uses ELF lexical agents to implement simple semantic orientation calculation. In Figure 1, each word is a lexical agent which has two features – score and total. All the lexical agents for actual words inherit from an abstract lexical agent node called lexical-agent$_1$. This nodes specifies a value for score of 0 (neutral) and a rules for calculating the total, by adding the score to 'prev total' – the total from the previous word. All the word nodes inherit both these specifications, except the word 'good' which specifies its own score of +3, overriding the (default) inheritance from lexical-agent$_1$. The resulting values for score and total are shown in the figure, and the SO score for the whole sentence can be read off from the value of the total feature for the last word.

In Figure 2, we extend this model with an another agent, lexical-agent$_2$, which describes a rule for intensifiers. This rule says that if the previous word is 'very', then this word's sentiment score has to be multiplied by a factor of 2. In this example, lexical-agent$_2$ is only used for sentiment-bearing words, such as 'good' – neutral words just inherit from lexical-agent$_1$ as before. Therefore, the sentiment score of 'good' changes and all other words' scores remain as before.

4 Modelling SO-CAL in Galadriel

In order to test out the Galadriel system architecture, we aimed to model SO-CAL in Galadriel. In this section we provide the key steps of the modelling process. We ended up creating a total of 6 models in Galadriel for SO-CAL features. Each model is used to capture one feature of SO-CAL. In Galadriel, we named the models as sent1, sent2 and so on.

4.1 Model sent1: Aggregating SO scores

We have four different dictionaries (used for SO-CAL) for the parts of speech adjectives, adverbs, nouns and verbs with their SO values (between +5 and-5). As discussed above, in order to get total SO value of document, SO-CAL aggregates the SO value of each word present in the document. In Galadriel, model sent1 is a simple model where each word has associated with it its own SO score and a total score for the document up to that point. This is as show in Figure 1, above, except that the SO scores come from the dictionaries, rather than being explicitly specified.

4.2 Model sent2: Intensification

Intensifiers do not contribute propositional meaning of a clause, and they generally do not have any sentiment of their own. But they give additional emotional context to a word they modify, which means intensifiers change the semantic intensity of that word. The words whose SO values are being modified by intensifiers are usually their neighbouring lexical item. Taboada et al. represented value of an intensifier as percentage, and these values are listed in the SO-CAL dictionaries. Figure 3 shows our modelling of intensifiers, which uses the same approach as in Figure 2, but allowing for different intensification factors (from the dictionaries), and making more explicit the inheritance between models sent2 and sent1.

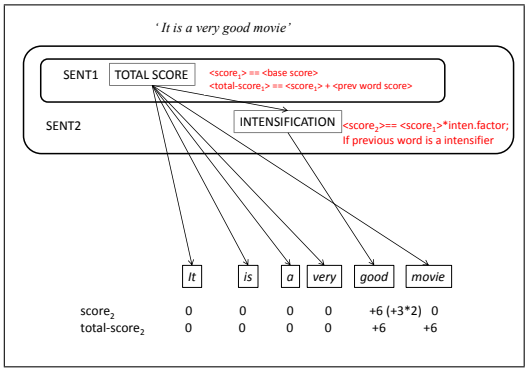

Figure 3 Model sent2 for intensifiers inheriting from model sent1.

4.3 Model sent3: Negation

Negation words usually reverse the opinion of a sentence. Two methods are applied for dealing with negators. They are the switch negation method, where the polarity of the lexical item next to negator will be switched, and the shift negation method, where the SO value of aword which needs to be negated is shifted towards the opposite polarity by a fixed amount. Negation words include 'not','never','no','nobody'. Similar to intensifiers, negators do not have SO values themselves and so are categorised as neutral. Taboata et al. defined any negator as negating the opinion expressed within the same clause. In order to identify a clause or sentence, a list of end punctuation words is created. This allows the identification of clauses and sentences in a document. Moreover, Taboada et al. argued that the switch negation does not work in certain cases. Therefore they implemented the shift negation method. They introduced a constant number 4 and instead of changing the sign they shifted SO value toward the opposite polarity by the constant 4.

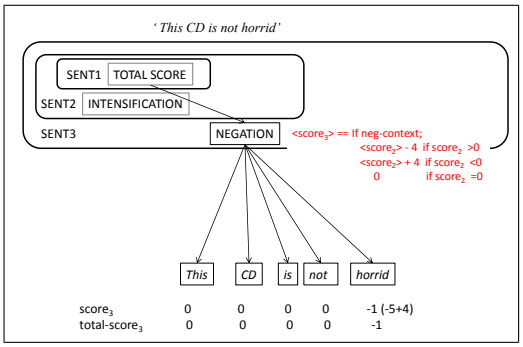

Figure 4 Model sent3 for negation:neg-context adjust the sentiment score.

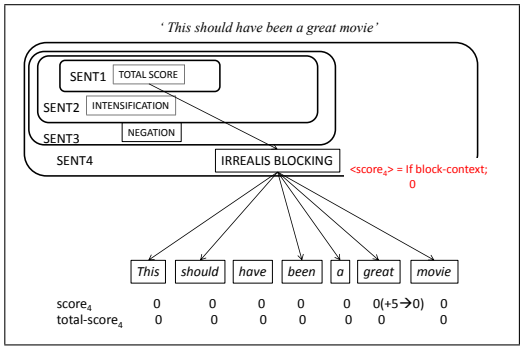

Figure 5 Model sent4:block-context changes sentiment scores to 0.

To model negation in Galadriel, first clauses and sentences are identified. Then any negation words within a clause/sentence negates opinion expressed within the same clause or sentence. In this model a new feature called 'neg-context' is introduced for each and every word in the document. The feature 'neg-context' takes the value either 'yes' or 'no'. Any word which could be negated by a negator, is assigned a 'neg-context' value as 'yes' otherwise 'no'. Finally, following SO-CAL, the shift negation rule is applied to the words which have a neg-context value of yes. (See Figure 4.)

4.4 Model sent4: Irrealis Blocking

Taboada et al. identified a number of irrealis markers which introduce non-factual context. Such markers indicate the words appearing in a clause/sentence are not reliable for the purpose of sentiment analysis. These words change the meaning of sentiment-bearing words and such words are named 'irrealis markers'. Their list of irrealis markers includes conditional markers ('if'), certain verbs, ('doubt', 'expect'), negative polarity items, words enclosed in quotes and questions.

In Galadriel, Taboada et al.'s list of irrealis marker is categorized under a hierarchical lexical node called 'mark'. To model SO-CAL's irrealis blocking feature in Galadriel, a new feature called 'block-context' with possible values 'yes' or 'no' is introduced. Similar to model sent3, the 'block-context' feature also uses end punctuation words to assign its own value, as irrealis blocking applies only within a clause or sentence. In addition, a 'ques-context' feature is used to decide whether the clause/sentence is a question. Then, if any determiners are found within the clause/sentence, irrealis blocking is ignored (see Figure 5).

4.5 Model sent5 and Model sent6: Text-Level Features

Taboada et al. believed lexicon-based sentiment classifiers generally favour positive language statements and so previous sentiment research shows a positive bias. Moreover they said, the repetition of a sentiment word found in a sentence shows sentiment depending on how many times the sentiment word is present in the sentence. SO-CAL may show strong positive sentiment, for example in Figure 6, due to the repetition of 'excellent' word. However, Taboada et al. suggested the number of appearance of a sentiment word in a sentence should not decide its overall sentiment intensity. In order to overcome above problems, firstly SO-CAL increased the final SO value of any negative expression by 50% . Secondly, they decreased the weight of words, which appear more often in the document. In this way, they

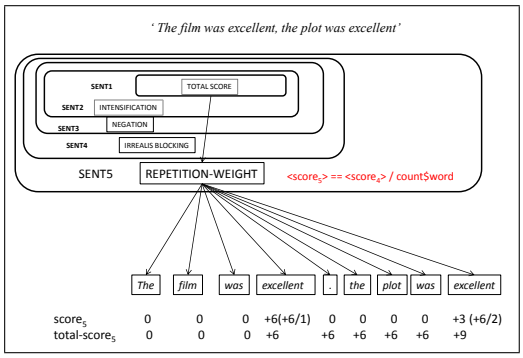

Figure 6 Model sent5: changes sentiment score of the word, dependent on its word count.

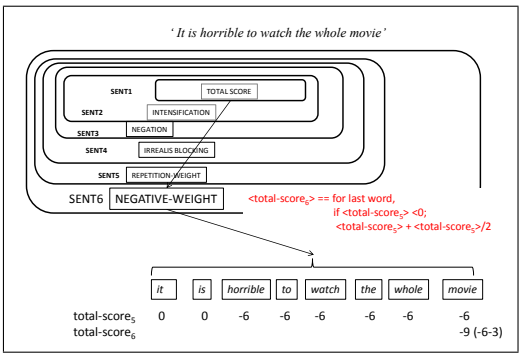

Figure 7 Model sent6: changes the total score, if it is negative.

Table 1 Performance of SO-CAL and Galadriel models for only adjective and all words.

	Only adjectives		All words		
Datasets	**SO-CAL**	**sent1**	**SO-CAL**	**sent1**	**sent6**
Epinions	72.25%	68.89%	80%	65.04%	60.68%
Movie reviews	76.63%	71%	76.37%	70%	65%

decided to override the SO value of the nth appearance of a word with 1/n of its full SO value.

To model SO-CAL's feature for repetition weight of words in Galadriel, a new feature called 'count $word' is introduced, where '$word' is a DATR variable, so this definition works for different actual words, for instance <count excellent>, <count horrid> . This feature allows us to count how many times a word is present in a document. Thus the sentiment score of the word ($word) is divided by 'count $word' to produce the final score for the word (see Figure 6). To model negation weighting, first the system decides whether the overall sentiment is negative. If so, the total score is increased by 50% (see figure 7).

5 Evaluation

We have collected the whole dataset and the dictionary used by SO-CAL. SO-CAL's dictionary contains list of words (adjectives, adverbs, nouns and verbs) with their SO (semantic orientation) values (between −5 and +5). In addition, it has a list of intensifiers with their values in factors (with plus and minus sign). We tested SO-CAL in Galadriel using two data sets, which were based on those used in [11]. The data sets are:

- **Epinions**: 50 reviews each of: books, cars, computers, cookware, hotels, movies, music and phones. As a first step of evaluation we used total 46 (24 positive and 22 negative) reviews.
- **Movies**: 1900 texts from the polarity data set [7]. We used 20 (10 positive and 10 negative) reviews.

We tested Galadriel in several configurations, simulating SO-CAL's 'only adjectives' and 'all words' (including sentiment for adverbs, nouns and verbs) settings, and for all six Galadriel models (sent1 – sent6). Table 1 shows the performances of SO-CAL and Galadriel with adjectives and all words in sent1 and sent6. Table 2 and Table 3 show performances of SO-CAL (all words) with different features and different models of Galadriel (all words)

Table 2 Performance of SO-CAL (all words) using various options.

Features	Epinions	Movies
simple	65.25%	68.05%
negation	67.75%	70.10%
neg+intensifiers	69.25%	73.47%
neg+inten+irrealis	78.25%	75.08%
neg+inten+irr+ neg weight	80.00%	76.37%
neg+inten+irr+ neg w+rep w	80.00%	76.37%

Table 3 Performance of Galadriel models (all words).

Models	Epinions	Movies
sent1	65.04%	70%
sent2	68.02%	72%
sent3	64.03%	69%
sent4	66.50%	67%
sent5	62.12%	67%
sent6	60.68%	65%

Table 4 Comparing performance of SO-CAL and Galadriel on positive and negative reviews.

Reviews	SO-CAL			Galadriel		
	Pos-F	Neg-F	Accuracy	Pos-F	Neg-F	Accuracy
Books	0.69	0.77	0.74	0.82	0.68	0.75
Cars	0.80	0.75	0.78	0.73	0.63	0.68
Computers	0.90	0.89	0.90	0.71	0.44	0.58
Cookware	0.79	0.76	0.78	0.82	0.25	0.54
Hotels	0.80	0.70	0.76	0.75	0.28	0.52
Movies	0.76	0.79	0.78	0.78	0.44	0.61
Music	0.83	0.81	0.82	0.75	0.33	0.54
Phones	0.85	0.83	0.84	0.75	0.66	0.71
Total	0.81	0.79	0.80	0.76	0.46	0.61

respectively. Table 4 shows comparison of the performance of SO-CAL and Galadriel across review types and on positive and negative reviews.

6 Discussion

In this paper, we have shown how the lexicon-based approach to sentiment analysis can be modelled by using inheritance based modelling techniques. Although we are not aiming to match performance of SO-CAL, we provided Galadriel performance figures in different experimental set-ups. We only aimed to show that SO-CAL features can be modelled in Galadriel.

We also have been modelling slightly different existing lexicon-based sentiment analysis approach [2] which is an aspect-based model in Galadriel and merging with SO-CAL, while identifying novel techniques. These models will be evaluated by comparing the existing original methods. From these analyses, an integrated inheritance model of sentiment knowledge of words will be identified and it will be extended to a model of sentiment analysis. In this way the entire sentiment analysis task will be coded as a 'lexical description' task.

We aim to introduce insights from other systems, in particular machine learning approaches, into model. To illustrate, we aim to use Galadriel to handle phrases that are commonly used to express sentiment. In order to handle such phrases, we will focus on building a model in Galadriel, using a corpus-based machine learning methodology to refine this model with examples derived from corpus data. This allows supporting exceptions to general rules.

Acknowledgements. I would like to thank Dr. Maite Taboada, who kindly gave me access to the SO-CAL system, its dictionaries and the datasets, and my supervisors Dr. Roger Evans and Dr. Gulden Uchyigit, for their support in producing this paper.

References

1 Walter Daelemans, Koenraad De Smedt, and Gerald Gazdar. Inheritance in natural language processing. *Computational Linguistics*, 18(2):205–218, 1992.

2 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.

3 Roger Evans. The extended lexicon: language processing as lexical description. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 270–276, Hissar, Bulgaria, September 2013. RANLP 2011 Organising Committee.

4 Roger Evans and Gerald Gazdar. DATR: A language for lexical knowledge representation. *Computational Linguistics*, 22(2):167–216, 1996.

5 Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125, 2006.

6 Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer, 2012.

7 Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. ACL, 2004.

8 Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

9 Livia Polanyi and Annie Zaenen. Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*, pages 1–10. Springer, 2006.

10 Maite Taboada, Caroline Anthony, and Kimberly Voll. Methods for creating semantic orientation dictionaries. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pages 427–432, 2006.

11 Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

12 Maite Taboada and Jack Grieve. Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Re# port SS# 04# 07), Stanford University, CA, pp. 158q161. AAAI Press*, 2004.

13 Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. ACL, 2002.