RA-SR: Using a ranking algorithm to automatically building resources for subjectivity analysis over annotated corpora

Yoan Gutiérrez, Andy González
University of Matanzas, Cuba
yoan.gutierrez@umcc.cu, andy.gonzalez@infonet.umcc.cu

Antonio Fernández Orquín, Andrés Montoyo, Rafael Muñoz
University of Alicante, Spain
antonybr@yahoo.com, {montoyo, rafael}@dlsi.ua.es

Abstract

In this paper we propose a method that uses corpora where phrases are annotated as Positive, Negative, Objective and Neutral, to achieve new sentiment resources involving words dictionaries with their associated polarity. Our method was created to build sentiment words inventories based on senti-semantic evidences obtained after exploring text with annotated sentiment polarity information. Through this process a graph-based algorithm is used to obtain auto-balanced values that characterize sentiment polarities well used on Sentiment Analysis tasks. To assessment effectiveness of the obtained resource, sentiment classification was made, achieving objective instances over 80%.

1 Introduction

In recent years, textual information has become one of the most important sources of knowledge to extract useful data. Texts can provide factual information, such as: descriptions, lists of characteristics, or even instructions to opinion-based information, which would include reviews, emotions or feelings. These facts have motivated dealing with the identification and extraction of opinions and sentiments in texts that require special attention. Among most widely used terms in Natural Language Processing, in concrete in Sentiment Analysis (SA) and Opinion Mining, is the subjectivity term proposed by (Wiebe, 1994). This author defines it as "linguistic expression of somebody's opinions, sentiments, emotions, evaluations, beliefs and speculations". Another important aspect opposed to subjectivity is the objectivity, which constitute a fact expression (Balahur, 2011). Other interesting terms also proposed by (Wiebe *et al.*, 2005) considers, private state, theses terms involve opinions, beliefs, thoughts, feelings, emotions, goals, evaluations and judgments.

Many researchers such as (Balahur *et al.*, 2010; Hatzivassiloglou *et al.*, 2000; Kim and Hovy, 2006; Wiebe *et al.*, 2005) and many others have been working in this way and related areas. To build systems able to lead SA challenges it is necessary to achieve sentiment resources previously developed. These resources could be annotated corpora, affective semantic structures, and sentiment dictionaries.

In this paper we propose a method that uses annotated corpora where phrases are annotated as Positive, Negative, Objective and Neutral, to achieve new resources for subjectivity analysis involving words dictionaries with their associated polarity.

The next section shows different sentiment and affective resources and their main characteristics. After that, our proposal is developed in section 3. Section 4, present a new sentiment resource obtained after evaluating RA-SR over many corpora. Section 5 described the evaluation and analysis of the obtained resource, and also an assessment of the obtained resource in Sentiment Classification task. Finally, conclusion and further works are presented in section 6.

2 Related work

It is known that the use of sentiment resources has proven to be a necessary step for training and evaluation for systems implementing sentiment analysis, including also fine-grained opinion mining (Balahur, 2011).

Different techniques have been used into product reviews to obtain lexicons of subjective words with their associated polarity. We can study the relevant research promoted by (Hu and Liu, 2004) which start with a set of seed adjectives ("good" and "bad") and reinforce the semantic knowledge applying a expanding the lexicon with synonymy and antonymy relations provided by WordNet (Miller *et al.*, 1990). As result of Hu and Liu researches an Opinion Lexicon is obtained with around 6800 positive

Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 94–99,
Atlanta, Georgia, 14 June 2013. ©2013 Association for Computational Linguistics

and negative English words (Hu and Liu, 2004; Liu *et al.*, 2005).

A similar approach has been used in building WordNet-Affect (Strapparava and Valitutti, 2004). In this case the building method starting from a larger of seed affective words set. These words are classified according to the six basic categories of emotion (joy, sadness, fear, surprise, anger and disgust), are also expanded increase the lexicon using paths in WordNet.

Other widely used in SA has been SentiWordNet resource (Esuli and Sebastiani, 2006)). The main idea that encouraged its construction has been that "terms with similar glosses in WordNet tend to have similar polarity".

Another popular lexicon is MicroWNOp (Cerini *et al.*, 2007). It contains opinion words with their associated polarity. It has been built on the basis of a set of terms extracted from the General Inquirer[1] (Stone *et al.*, 1996).

The problem is that these resources do not consider the context in which the words appear. Some methods tried to overcome this critique and built sentiment lexicons using the local context of words.

We can mentioned to (Pang *et al.*, 2002) whom built a lexicon with associated polarity value, starting with a set of classified seed adjectives and using conjunctions ("and") disjunctions ("or", "but") to deduce orientation of new words in a corpus.

(Turney, 2002) classifies words according to their polarity based on the idea that terms with similar orientation tend to co-occur in documents.

On the contrary in (Balahur and Montoyo, 2008b), is computed the polarity of new words using "polarity anchors" (words whose polarity is known beforehand) and Normalized Google Distance (Cilibrasi and Vitányi, 2007) scores using as training examples opinion words extracted from "pros and cons reviews" from the same domain. This research achieved the lexical resource Emotion Triggers (Balahur and Montoyo, 2008a).

Another approach that uses the polarity of the local context for computing word polarity is the one presented by (Popescu and Etzioni, 2005), who use a weighting function of the words around the context to be classified.

All described resources have been obtained manually or semi-automatically. Therefore, we focus our target in archiving automatically new sentiment resources supported over some of aforementioned resources. In particular, we will offer contributions related with methods to build sentiment lexicons using the local context of words.

3 Our method

We propose a method named RA-SR (using Ranking Algorithms to build Sentiment Resources) to build sentiment words inventories based on senti-semantic evidences obtained after exploring text with annotated sentiment polarity information. Through this process a graph-based algorithm is used to obtain auto-balanced values that characterize sentiment polarities widely used on Sentiment Analysis tasks. This method consists of three main stages: **(I)** Building contextual words graphs; **(II)** Applying ranking algorithm; and **(III)** Adjusting sentiment polarity values.

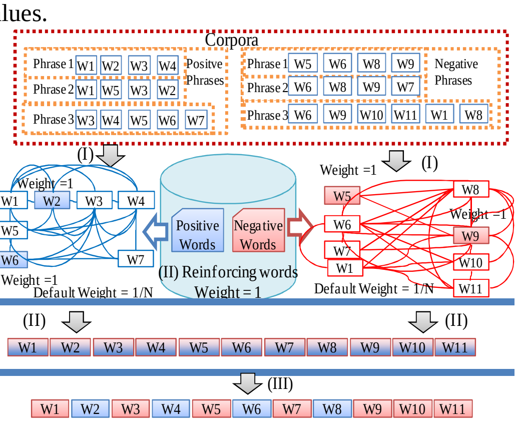

Figure 1. Resource walkthrough development process.

These stages are represented in the diagram of Figure 1, where the development process begins introducing two corpuses of annotated sentences with positive and negative sentences respectively. Initially, a preprocessing of the text is made applying Freeling pos-tagger (Atserias *et al.*, 2006) version 2.2 to convert all words to lemmas[2]. After that, all lemmas lists obtained are introduced in RA-SR, divided in two groups (i.e. positive and negative candidates, $Spos$ and $Sneg$).

3.1 Building contextual words graphs

Giving two sets of sentences ($Spos$ and $Sneg$) annotated as positive and negative respectively, where $Spos = [L_{pos1}, \dots, L_{posM}]$ and $Sneg = [L_{neg1}, \dots, L_{negM}]$ contains list L involving words lemmatized by Freeling 2.2 Pos-Tagger

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

[2] Lemma denotes canonic form of the words.

(Atserias *et al.*, 2006), a process to build two lexical contextual graphs, $Gpos$ and $Gneg$ is applied. Those sentences are manually annotated as positive and negative respectively. These graphs involve lemmas from the positive and negative sentences respectively.

A contextual graph G is defined as an undirected graph $G = (V, E)$, where V denotes the set of vertices and E the set of edges. Given the list $L = [l_1 \dots l_N]$ a lemma graph is created establishing links among all lemmas of each sentence, where words involved allow to interconnect sentences l_i in G. As a result word/lemma networks $Gpos$ and $Gneg$ are obtained, where $L = V = [l_1 \dots l_N]$ and for every edge $(l_i, l_j) \in E$ being $l_i, l_j \in V$. Therefore, l_i and v_i are the same.

Then, having two graphs, we proceed to initialize weight to apply graph-based ranking techniques in order to auto-balance the particular importance of each v_i into $Gpos$ and $Gneg$.

3.2 Applying ranking algorithm

To apply a graph-based ranking process, it is necessary to assign weights to the vertices of the graph. Words involved into $Gpos$ and $Gneg$ take the default value $1/\mathrm{N}$ as their weight to define the weight of v vector, which is used in our proposed ranking algorithm. In the case where words are identified on the sentiment repositories (see Table 2) as positive or negative, in relation to their respective graph, a weight value of 1 (in a range $[0 \dots 1]$) is assigned. N represents the maximum quantity of words in the current graph. Thereafter, a graph-based ranking algorithm is applied in order to structurally raise the graph vertexes' voting power. Once the reinforcement values are applied, the proposed ranking algorithm is able to increase the significance of the words related to these empowered vertices.

The PageRank (Brin and Page, 1998) adaptation, which was popularized by (Agirre and Soroa, 2009) in Word Sense Disambiguation thematic, and the one that has obtained relevant results, was an inspiration to us in this work. The main idea behind this algorithm is that, for each edge between v_i and v_j in graph G, a vote is made from v_i to v_j. As a result, the relevance of v_j is increased.

On top of that, the vote strength from i to j depends on $v_i's$ relevance. The philosophy behind it is that, the more important the vertex is, the more strength the voter would have. Thus, PageRank is generated by applying a random walkthrough from the internal interconnection of G, where the final relevance of v_i represents the random walkthrough probability over G, and ending on v_i.

In our system, we apply the following equation and configuration:

$$\mathbf{Pr} = cM\mathbf{Pr} + (1 - c)v \quad (1)$$

Where: M is a probabilistic transition matrix $N \times N$, being $M_{j,i} = \frac{1}{d_i}$ if a link from v_i to v_j exist, in other case zero is assigned; v is a vector $N \times 1$ with values previously described in this section; **Pr** is the probabilistic structural vector obtained after a random walkthrough to arrive to any vertex; c is a dumping factor with value 0.85, and like in (Agirre and Soroa, 2009) we used 30 iterations.

A detailed explanation about the PageRank algorithm can be found in (Agirre and Soroa, 2009).

After applying PageRank, in order to obtain standardized values for both graphs, we normalize the rank values by applying the following equation:

$$\mathbf{Pr}_i = \mathbf{Pr}_i / Max(\mathbf{Pr}) \quad (2)$$

Where $Max(\mathbf{Pr})$ obtains the maximum rank value of $\boldsymbol{Pr}$ vector.

3.3 Adjusting sentiment polarity values

After applying the PageRank algorithm on $Gpos$ and $Gneg$, and having normalized their ranks, we proceed to obtain a final list of lemmas (named Lf) while avoiding repeated elements. Lf is represented by Lf_i lemmas, which would have, at that time, two assigned values: Positive, and Negative, which correspond to a calculated rank obtained by the PageRank algorithm.

At that point, for each lemma from Lf, the following equations are applied in order to select the definitive subjectivity polarity for each one:

$$Pos = \begin{cases} Pos - Neg\, ; \; Pos > Neg \\ 0 \qquad\qquad ; otherwise \end{cases} \quad (3)$$

$$Neg = \begin{cases} Neg - Pos\, ; \; Neg > Pos \\ 0 \qquad\qquad ; otherwise \end{cases} \quad (4)$$

Where Pos is the Positive value and Neg the Negative value related to each lemma in Lf.

In order to standardize the Pos and Neg values again and making them more representative in a [0…1] scale, we proceed to apply a normalization process over the Pos and Neg values.

Following and based on the objective features commented by (Baccianella *et al.*, 2010), we assume their same premise to establish objective values of the lemmas. Equation (5) is used to this proceeding, where Obj represent the objective value.

$$Obj = 1 - |Pos - Neg| \quad (5)$$

4 Sentiment Resource obtained

At the same time we have obtained a Lf where each word is represented by Pos, Neg and Obj values, acquired automatically from annotated sentiment corpora. With our proposal we have been able to discover new sentiment words in concordance of contexts in which the words appear. Note that the new obtained resource involves all lemmas identified into the annotated corpora. Pos, Neg, and Obj are nominal values between range [0 ... 1].

5 Evaluation

In the construction of the sentiment resource we used the annotated sentences provided from corpora described on Table 1. Note that we only used the sentences annotated positively and negatively. The resources involved into this table were a selection made to prove the functionality of the words annotation proposal of subjectivity and objectivity.

The sentiment lexicons used were provided from WordNetAffect_Categories[3] and opinion-words[4] files and shown in detail in Table 2.

Corpus	Neg	Pos	Obj	Neu	Obj or Neu	Unknow	Total
computational-intelligence[5]	6982	6172	-	-	-	-	13154
tweeti-b-sub.dist_out.tsv[6]	176	368	110	34	-	-	688
b1_tweeti-objorneu-b.dist_out.tsv[6]	828	1972	788	1114	1045	-	5747
stno[7]	1286	660		384	-	10000	12330
Total	9272	9172	898	1532	1045	10000	31919

Table 1. Corpora used to apply RA-SR.

Sources	Pos	Neg	Total
WordNet-Affects_Categories (Strapparava and Valitutti, 2004)	629	907	1536
opinion-words (Hu and Liu, 2004; Liu *et al.*, 2005)	2006	4783	6789
Total	2635	5690	8325

Table 2. Sentiment Lexicons.

Some issues were taking into account through this process. For example, after obtaining a contextual graph G factotum words are present in mostly of the involved sentences (i.e. verb "*to be*"). This aspect is very dangerous after applying PageRank algorithm, because this algorithm because this algorithm strengthens the nodes possessing many linked elements. For that reason, the subtractions $Pos - Neg$ and $Neg - Pos$ are applied, where the most frequently words in all contexts obtains high values and being the subtraction a damping factor.

Following an example; when we take the verb "*to be*", before applying equation (2), verb "*to be*" archives the highest values into each context graph ($Gpos$ and $Gneg$), 9.94 and 18.67 rank values respectively. These values, applying equation (2), are normalized obtaining both $Pos = 1$ and $Neg = 1$ in a range [0...1]. Finally, when the next steps are executed (Equations (3) and (4)) verb "*to be*" achieves $Pos = 0$, $Neg = 0$ and therefore $Obj = 1$. Through this example it seems as we subjectively discarded words that appear frequently in both contexts (Positive and Negative contexts).

Using the corpora from Table 1 we obtain 25792 sentimentally annotated lemmas with Pos, Neg and Obj features. Of them 12420 positive and 11999 negative lemmas were discovered, , and 1373 words already derived from existing lexical resources.

Another contribution has been the Pos, Neg and Obj scores assigned to words of lexical inventory, which were used to reinforce the contextual graphs in the building process. Those words in concordance to our scenario count 842 Positives and 383 Negatives.

5.1 Sentiment Resource Applied on Sentiment Analysis

To know if our method offers resources that improve the SA state of the art, we propose a **baseline** supported on the sentiment dictionaries, and other method (Ranking Sentiment Resource (**RSR**)) supported over our obtained resource. The **baseline** consists on analyzing sentences applying Equation (6) and Equation (7).

$$PosMeasure = \frac{PosCount}{WordCount} \quad (6)$$

$$NegMeasure = \frac{NegCount}{WordCount} \quad (7)$$

Where: $PosCount$ is the total of positive words (aligned with the sentiment dictionaries) in the sentence; $NegCount$ is the total of negative words (aligned with the sentiment dictionaries)

[3] http://wndomains.fbk.eu/wnaffect.html
[4] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
[5] A sentimental corpus obtained applying techniques developed by GPLSI department. See (http://gplsi.dlsi.ua.es/gplsi11/allresourcespanel)
[6] Train dataset of Semeval-2013 (Task 2. Sentiment Analysis in Twitter, subtask b.)
[7] Test dataset of NTCIR Multilingual Opinion Analysis Task (MOAT) http://research.nii.ac.jp/ntcir/ntcir-ws8/meeting/

in the sentence; $WordCount$ is the total of words in the sentence.

Using these measures over the analyzed sentences, for each sentence, we obtain two attributes, $PosMeasure$ and $NegMeasure$; and a third attribute (named Classification) corresponding to its classification.

On the other hand, we propose **RSR**. This SA method uses in a different way the Equation (6) and Equation (7), and introduces Equation (8).

$$ObjMeasure = \frac{ObjCount}{WordCount} \quad (8)$$

Being $PosCount$ the sum of Positive ranking values of the sentence words, aligned with the obtained resource (Lf); $NegCount$ the sum of Negative ranking values of the sentence words, aligned with the obtained resource (Lf); and $ObjCount$ the sum of Objective ranking values of the sentence words, aligned with the obtained resource (Lf).

In RSR method we proved with two approach, RSR ($1/d_i$) and RSR ($1-(1/d_i)$). The first approach is based on a resource developed using PageRank with $M_{j,i} = 1/d_i$ and the other approach is using $M_{j,i} = 1 - (1/d_i)$. Table 3 shows experimentation results.

The evaluation has been applied over a corpus provided by “Task 2. Sentiment Analysis in Twitter, subtask b”, in particular tweeti-b-sub.dist_out.tsv file. This corpus contains 597 annotated phrases, of them Positives (314), Negatives (155), Objectives (98) or Neutrals (30). For our understanding this quantity of instances offers a representative perception of RA-SR contribution; however we will think to evaluate RA-SR over other corpora in further researches.

	C	I	R. Pos (%)	R. Neg (%)	R. Obj (%)	R. Neu (%)	Total P. (%)	Total R. (%)
Baseline	366	231	91.1	51.6	0.0	0.0	48.2	61.3
RSR($1/d_i$)	416	181	87.3	39.4	80.6	6.7%	67.8	69.7
RSR($1-(1/d_i$)	469	128	88.5	70.3	81.6	6.7%	76.8	78.6

Table 3. Logistic function (Cross-validation 10 folds) over tweeti-b-sub.dist_out.tsv[8] corpus (597 instances). Recall (R), Precision (P), Correct (C), Incorrect (I).

As we can see the baseline only is able to dealing with negative and positive instances. Is important to remark that our proposal starting up knowing only the words used in baseline and is able to growing sentiment information to other words related to them. We can see this fact on Table 3, RSR is able to classify objective instances over 80% of Recall and the baseline does not.

Other relevant element is the recall difference between RSR ($1/d_i$) and RSR $(1-(1/d_i)$. Traditionally $(1/d_i)$ result value has been assigned to M in PageRank algorithm. We have demonstrated that in lexical contexts RSR ($1-(1/d_i)$) approach offers a better performance of PageRank algorithm, showing recall differences around 10 perceptual points.

6 Conclusion and further works

As a conclusion we can say that our proposal is able to automatically increase sentiment information, obtaining 25792 sentimentally annotated lemmas with Pos, Neg and Obj features. Of them 12420 positive and 11999 negative lemmas were discovered.

In other hand, The RSR is capable to classify objective instances over 80% and negatives over 70%. We cannot tackle efficiently neutral instances, perhaps it is due to the lack of neutral information in the sentiment resource we used. Also, it could be due to the low quantity of neutral instances in the evaluated corpus.

In further research we will evaluate RA-SR over different corpora, and we are also going to deal with the number of neutral instances.

The variant RSR$(1-(1/d_i)$ performs better than RSR$(1/d_i)$ one. This demonstrates that in lexical contexts using PageRank with $M_{j,i} = 1 - (1/d_i)$ offers a better performance. Other further work consists in exploring Social Medias to expand our retrieved sentiment resource obtaining real time evidences that occur in Web 2.0.

Acknowledgments

This research work has been partially funded by the Spanish Government through the project TEXT-MESS 2.0 (TIN2009-13391-C04), "Análisis de Tendencias Mediante Técnicas de Opinión Semántica" (TIN2012-38536-C03-03) and “Técnicas de Deconstrucción en la Tecnologías del Lenguaje Humano” (TIN2012-31224); and by the Valencian Government through the project PROMETEO (PROMETEO/2009/199).

References

Agirre, E. and A. Soroa. Personalizing PageRank for Word Sense Disambiguation. Proceedings of the 12th conference of the European chapter of the

[8] Semeval-2013 (Task 2. Sentiment Analysis in Twitter, subtask b.)

Association for Computational Linguistics (EACL-2009), Athens, Greece, 2009. p.

Atserias, J.; B. Casas; E. Comelles; M. González; L. Padró and M. Padró. FreeLing 1.3: Syntactic and semantic services in an opensource NLP library. Proceedings of LREC'06, Genoa, Italy, 2006. p.

Baccianella, S.; A. Esuli and F. Sebastiani. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. 7th Language Resources and Evaluation Conference, Valletta, MALTA., 2010. 2200-2204 p.

Balahur, A. Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types. Department of Software and Computing Systems. Alacant, Univeristy of Alacant, 2011. 299. p.

Balahur, A.; E. Boldrini; A. Montoyo and P. Martinez-Barco. The OpAL System at NTCIR 8 MOAT. Proceedings of NTCIR-8 Workshop Meeting, Tokyo, Japan., 2010. 241-245 p.

Balahur, A. and A. Montoyo. Applying a culture dependent emotion trigger database for text valence and emotion classification. Procesamiento del Lenguaje Natural, 2008a. p.

Balahur, A. and A. Montoyo. Building a recommender system using community level social filtering. 5th International Workshop on Natural Language and Cognitive Science (NLPCS), 2008b. 32-41 p.

Brin, S. and L. Page The anatomy of a large-scale hypertextual Web search engine Computer Networks and ISDN Systems, 1998, 30(1-7): 107-117.

Cerini, S.; V. Compagnoni; A. Demontis; M. Formentelli and G. Gandini Language resources and linguistic theory: Typology, second language acquisition, English linguistics (Forthcoming), chapter Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining., 2007.

Cilibrasi, R. L. and P. M. B. Vitányi The Google Similarity Distance IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2007, VOL. 19, NO 3.

Esuli, A. and F. Sebastiani. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. Fifth international conference on Languaje Resources and Evaluation Genoa - ITaly., 2006. 417-422 p.

Hatzivassiloglou; Vasileios and J. Wiebe. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. International Conference on Computational Linguistics (COLING-2000), 2000. p.

Hu, M. and B. Liu. Mining and Summarizing Customer Reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), USA, 2004. p.

Kim, S.-M. and E. Hovy. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In Proceedings of workshop on sentiment and subjectivity in text at proceedings of the 21st international conference on computational linguistics/the 44th annual meeting of the association for computational linguistics (COLING/ACL 2006), Sydney, Australia, 2006. 1-8 p.

Liu, B.; M. Hu and J. Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. Proceedings of the 14th International World Wide Web conference (WWW-2005), Japan, 2005. p.

Miller, G. A.; R. Beckwith; C. Fellbaum; D. Gross and K. Miller Introduction to WordNet: An On-line Lexical Database International Journal of Lexicography, 3(4):235-244., 1990.

Pang, B.; L. Lee and S. Vaithyanathan. Thumbs up? Sentiment Classification using machine learning techniquies. EMNLP -02, the Conference on Empirical Methods in Natural Language Processing, USA, 2002. 79-86 p.

Popescu, A. M. and O. Etzioni. Extracting product features and opinions from reviews. Proccedings of HLT-EMNLP, Canada, 2005. p.

Stone, P.; D. C.Dumphy; M. S. Smith and D. M. Ogilvie The General Inquirer: A Computer Approach to Content Analysis The MIT Press, 1996.

Strapparava, C. and A. Valitutti. WordNet-Affect: an affective extension of WordNet. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, 2004. 1083-1086 p.

Turney, P. D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceeding 40th Annual Meeting of the Association for Computational Linguistic. ACL 2002, USA, 2002. 417-424 p.

Wiebe, J. Tracking point of view in narrative Computational Linguistic, 1994, 20(2): 233-287.

Wiebe, J.; T. Wilson and C. Cardie. Annotating Expressions of Opinions and Emotions in Language. Kluwer Academic Publishers, Netherlands, 2005. p.