Opinion Analysis from the Social Web Contributions

Kristína Machová

Dept. of Cybernetics and Artificial Intelligence, Technical University, Letná 9,
042 00, Košice, Slovakia,
kristina.machova@tuke.sk

Abstract. The paper focuses on the automatic opinion analysis related to web discussions. It introduces a method for solving basic problems of opinion analysis (determination of word subjectivity, polarity as well as intensity of this polarity). The method solves the reversion of polarity by negation as well as determination of polarity intensity of word combinations. A dynamic coefficient for the word combinations processing is introduced and an implementation of the method is presented. In addition, the paper describes test results of the presented implementation and discussion of these results as well.

Keywords: opinion analysis, opinion classification, social web, sentiment analysis, web discussions, dynamic coefficient

1 Introduction

Opinion analysis represents a domain, which is a firm part of the field of social web analysis. The social web can be considered as an upgrade of the classic web. The classic web can be illustrated with an idea of a world-wide billboard - anybody can publish some information piece or make it accessible for public inspection on the billboard (anybody who has necessary skills in web page creation - but considerably greater amount of web users have abilities only for reading this published information). On the other hand, the social web or web 2.0 reinforces social interactions among user and provides an opportunity for great majority of web users to contribute to web content. It can be said, that it increases the number of web content providers. Social interactions among users are enabled by communication within social nets, by the possibility to contribute to web discussions, and so on.

Specifically, discussion forums are large-scale data bases of opinions, attitudes and feelings of web users, who use the web for communication. Unlike classic data bases, they do not contain data in a structured form. For this reason, they need special methods for processing. One of such special methods is also opinion analysis. The main objective of opinion analysis is to summarize attitude of particular subscribers to some particular theme. This theme can be, for example, an evaluation of some product, political situation, person (e.g. active in politics), event or company.

Opinion analysis or opinion classification can be used in those fields where the aggregation of a large amount of opinions into integrated information is needed. The input to opinion classification can be represented by a large amount of discussion contributions (e.g. content of a discussion forum) and the output of the classification is summarising information, for example "Users are satisfied with this product" or "People perceive this reform negatively". From the point of view of a consumer, two kinds of information are important for decision making about purchase of a product. First, it is information about price and properties of the product, which usually are available on web pages of a producer or a seller. Second, it is information about satisfaction of other consumers with the product. The opinion classification can offer this information to prospective consumer. From the point of view of a producer, information about satisfaction and needs of consumers is also very important. The classic way of obtaining this information is performing market research. The market research carried out by telephone or by questionnaires is usually rather expensive and time consuming. The promptness of such information elicitation is a matter of principle. User contribution analysis provided by a system utilising opinion classification can offer the information about clients' satisfaction more quickly.

2 Related Works

Sometimes, the introduced opinion analysis is denoted as opinion mining, because it focuses on the extraction of positive or negative attitude of a participant to commented objects with the aid of mining techniques applied to text documents. Opinion mining can be extended from the level of whole texts perception to the level of extraction of properties of those objects which match users' interests [3]. Parallel approach to opinion mining is sentiment analysis [8]. Different access to web discussion processing is represented by the estimation of authority degree of some information sources, for example of actors contributing to discussion forums or social nets. An important technique for authoritative actors searching is visualization approach, which is introduced in [4]. Some effort was spent on semantically enrich algorithms for analysis of web discussion contributions [6].

Nowadays, opinion analysis has become an important part of social networks analysis. Existing opinion analysis systems use large vocabularies for opinion classification into positive or negative answer categories. Such approach was used in [2]. Authors studied accuracy of the opinion analysis of Spanish documents originated in the field of economic. This approach uses a regression model for classification into negative or positive opinions. Authors studied how quality depends on the granularity of opinions and rules, which were used in the regression model. Another study [5] was focused on the possibility of using lesser granularity without any significant precision decrease. The results of this study show no substantial difference between one and two parameter regression models as well as no statistically significant difference between models with different granularity. Thus, for example, simpler models can be used with the used sentiment scale reduced to five degrees only.

The presented approach uses a scale with five degrees for opinion classification as well, but it differs from the previous approaches in vocabulary cardinality. Our work focuses on creating vocabularies with strong orientation on the discussion domain, not so large but created directly from live discussions. We do not use regression models. First, words from discussions are classified into predefined categories and after that, this classification is transformed into another one enabling classification of the whole contribution into one of five degrees (strong negative, negative, neutral, positive and strong positive).

3 Basic Problems of Opinion Analysis

Three basic problems of opinion analysis are: *word subjectivity identification, word polarity (orientation) determination and determination of intensity of the polarity*. Opinion analysis focuses on those words, which are able to express *subjectivity* very well - mainly adjectives (e.g. ‘perfect’) and adverbs (e.g. ‘beautifully’) are considered. On the other hand, other word classes must be considered as well in order to achieve satisfactory precision, for example nouns (e.g. ‘bomb’) or verbs (e.g. ‘devastate’). The words with subjectivity are important for opinion analysis; therefore they are identified and inserted into the vocabulary. Words with subjectivity are inserted into the constructed vocabulary together with their polarity.

The *polarity of words* forms a basis for the polarity determination of the whole discussion. There are three basic degrees of polarity being distinguished: positive (e.g. ‘perfect’, ‘attract’), negative (e.g. ‘junk’, ‘shocking’, ‘absurdity’, ‘destroyed’) and neutral (e.g. ‘averaged’, ‘effectively’). This scale can be refined to use more possible levels if needed. The determination of the polarity of words is connected with a problem of word polarity reversion – the reversion can be done by using negation, for example ‘It was not very attractive film’. This problem serves as an argument for the extension of single words polarity determination to polarity determination of word combinations (considering whole sentences or parts of sentences).

The *intensity of word polarity* represents a measure of the ability of words to support the proof or disproof of a certain opinion. The polarity intensity of words can be determined according to a defined scale, which helps to classify words into more categories. Table 1 illustrates three such scales with different numbers of degrees.

Table 1. Scales using verbal or numerical representation of the intensity of word polarity

Number of Degrees	Scales of polarity intensity	
2	negative	Positive
6	weak, gently, strong negative	weak, gently, strong positive
8	-1, -2, -3, -4	1, 2, 3, 4

The polarity intensity can be expressed both verbally as well as numerically. The numerical representation is more suitable for subsequent processing by computers. Discussion contributions very often contain some word combinations, which increase (decrease) the weak (strong) intensity of polarity of an original word, for example: ‘surprisingly nice’, ‘high quality’, ‘markedly weaker’ and ‘extremely low-class’.

3.1 Classification Vocabulary Creation

In order to support the process of opinion analysis, it is necessary to create a vocabulary. The opinion analysis systems commonly utilise large vocabularies, which are called seed-lists. For example WorldNet can be used as a basis for the creation of such seed-list vocabulary. In accordance with [1], it is possible to derive tagsonomies from crowd. Similarly, we attempted to derive a vocabulary directly from web discussions. This vocabulary is specialized for a particular domain, the utilised web discussions focus on. Since it is possible to use this vocabulary for classification of words into predefined categories, we denote it as a classification vocabulary.

Many of web discussion respondents do use literary language far from perfectly. Therefore our system of opinion classification has to be able of the adaptation to colloquial language of users of the Internet including slang, absence of diacritical marks, and frequent mistakes.

4 Design of Opinion Classification Method

The design of an opinion classification method has to consider all steps of the classification process and provide them in the right and logical sequence. The method we have designed solves the following problems:

- Basic problems of opinion analysis
- Word polarity reversion by negation
- Determination of the intensity of polarity
- Establishment of a dynamic coefficient
- Polarity determination of word combinations

Our access takes into account not only polarity of single words but also the intensity of polarity of word combinations including negation. Our method analyzes texts of discussion contributions from a certain domain and for this domain a classification vocabulary is generated from the given texts. The quality of the vocabulary and its cardinality play the key role in the process of opinion classification.

The method transforms textual content of a discussion contribution into an array of words. Each word with subjectivity is assigned a numerical value (numerical code) as it is illustrated in Fig. 1. This value represents the category of word polarity to which the given word belongs (see Table 2). Particular sentences are identified. First non zero value of word category starts the creation of word combination procedure. The length of a certain combination is limited by a coefficient *K*. Each combination of words is also assigned a numerical value which represents a polarity degree from the <-3, 3> interval. The polarity degrees of all word combinations within the given text form the polarity of this text as a whole. Subsequently, the polarity of the whole discussion can be calculated from the polarities of all contributions (texts).

The whole contribution or discussion is considered to be positive/negative when it contains more positive/negative word combinations or contributions. The neutral contribution (discussion) contains the same number of positive and negative word combinations (contributions). This approach to neutrality determination is rather strict. A more benevolent approach uses the following rule for neutrality detection:

$$IF\ |Number\_pozit - Number\_negat| \leq\ H\ THEN\ neutrality$$

where threshold *H* represents range of neutrality, which can be changed by setting another value of the *H* parameter ($H \geq\ 1$ and it is an integer). Strict approach to neutrality with $H=0$ is more suitable for very short contributions, because such short contributions can contain only one sentence and only one positive or negative word. Wider neutrality range could absorb this word and subsequently the system of opinion classification can evaluate it as a neutral contribution. The wider neutrality range is more suitable for longer contributions processing.

4.1 Basic Problems Solution

In our approach, words with subjectivity are selected and the category value from a given scale (the scale from 0 to 9 is used) is assigned to each of these words, what is illustrated in Fig. 1. The words with positive polarity are classified to categories 1 or 8 (see Table 2). Similarly, the words representing negative polarity are classified to categories 2 or 9 and words with neutral polarity to 0 category.

Table 2. Categories of words polarity

weak positive and strong positive	1 and 8
weak negative and strong negative	2 and 9
Neutral	0
negation – polarity reversion	3
increasing of polarity intensity	4

To illustrate usage of these categories, Fig. 1 illustrates categorization of words into polarity categories based on the example ‘This mobile is marvellous and its functioning is reliable’. This word classification system also solves determination of the intensity of the polarity, because values 1 and 2 represent weak polarity in contrast to values 8 and 9, which represent strong polarity (being positive or negative). Thus, the designed method uses a five degree scale of the intensity of polarity determination (including neutral).

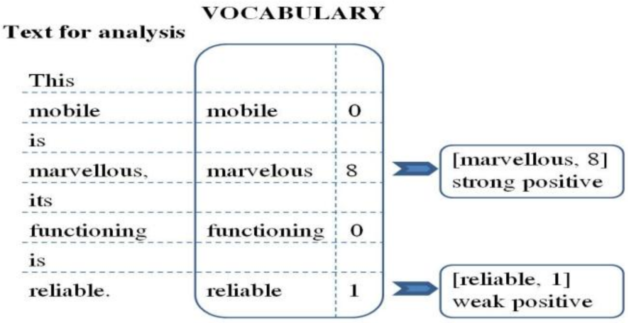

Fig. 1. Polarity determination of words from the sentence ‘This mobile is marvellous, its functioning is reliable.’

There is one more addition in our design for determining the intensity of polarity. It is the category 4 used for each word, which increases the intensity of polarity of another word in the same word combination (e.g. ‘high quality’).

To summarise, the used polarity categories are introduced in Table 2. All words with subjectivity are expected to be inserted into the classification vocabulary together with their category codes.

4.2 Word Polarity Reversion by Negation

The reversion of word polarity caused by the usage of negation enables to reflect actual meaning and therefore to increase precision of opinion classification. The words, which represent negation (e.g. 'none', 'no') belong to the category 3. This category can be used only in the combination with another category (1, 2, 8 or 9). It changes positive polarity into negative polarity and vice versa within the same degree of intensity (weak or strong) as it can be seen in Table 3.

Table 3. Word polarity reversion by negation

3 + 1	3 + 8	3 + 2	3 + 9
negation + weak positive = *weak negative*	negation + strong positive = *strong negative*	negation + weak negative = *weak positive*	negation + strong negative = *strong positive*

The polarity reversion is a rather complicated issue due to the fact, that the structure of various sentences is not homogenous. For example, the sentence 'This mobile isn't reliable' can be represented by the code 0031 (the code of a sentence is created by replacing each word with the number indicating its category). Another sentence 'It isn't, according to my opinion, reliable mobile' has the same meaning but different code 03000010. The aim of our technique is to recognise various codes 0031 and 03000010 as opinions with the same polarity. Thus, there is a need of some dynamic coefficient, which enables to estimate an appropriate length of those word combinations, which will be processed together as one lexical unit. In general, it enables to process one sentence as two different combinations – lexical units.

4.3 Determination of the Intensity of Polarity

Words, which increase the intensity of polarity, have no polarity and their influence on polarity of a lexical unit can be evaluated only within a combination with the given lexical unit. These words belong to the category 4. Table 4 presents two different examples of such combinations.

Table 4. Analysis of lexical units with word increasing polarity intensity

This	**mobile**	**is**	**totally**	**conforming**
0-neutral	0-neutral	0-neutral	**4 + intensity**	1-weak positive
It	**really**	**drives**	**me**	**mad**
0-neutral	**4 + intensity**	0-neutral	0-neutral	2-weak negative

Both these combinations contain a word increasing the intensity of polarity. The word combinations are represented with codes 00041 and 04002. Words from the category 4 are usually adverbs (e.g. 'very', 'really', 'totally'). Processing of the words enabling to increase the intensity of word polarity needs to use the dynamic coefficient in a similar manner as the negation processing.

4.4 Dynamic Word Combination Length

The designed method of opinion classification has an ambition to manage the variability of sentence structures using the dynamic coefficient *K*. The value of this parameter is being dynamically changed during processing of different lexical units. The dynamic coefficient adapts itself to the code length of a lexical unit (sequence of words) under investigation. The value *K* represents the number of words, which are included into the same word combination (beginning from the first non-zero word code in the sequence of words). In the case, when the value is higher than the number of words in the sentence, this value is dynamically decreased in order to ensure, that the combination contains only words from the investigated sentence, not from the beginning of the following sentence. A word combination can be shortened also in some other cases. For example, let us take the case *K=4* while the combination 3011 is being processed. In this case, two disjunctive combinations are created 301 (*K=3*) and 1 (*K=1*). On the other hand, the value can be increased in some cases. Table 5 illustrates the principle of using the dynamical coefficient.

Table 5. Principle of employing the dynamical coefficient *K* (Words processed within one combination are given in bold.)

K	Never	buy	this	nice	mobile
1	**3**	0	0	**1**	0
2	**3**	**0**	0	**1**	**0**
4	**3**	**0**	**0**	**1**	0

As we can see in Table 5, value *K=1* is not appropriate for processing of the sentence ‘Never buy this nice mobile!’, because negation ‘never’ would be in a combination different from the combination comprising the word ‘nice’, to which the negation is related. Setting *K=1* represents processing of words in isolation from each other. The alternative *K=2* allows processing of neighbouring words as combinations, but it does not prevent the isolation of negation from relating word either. This sentence can be satisfactorily processed only when the coefficient has value $K\geq 4$.

4.5 Polarity Determination of Word Combinations

Generation of suitable word combinations using the dynamic coefficient *K* is the key factor of effective opinion classification. These combinations are sets words (their cardinality differs according to changing value of *K*), to which a polarity degree, representing the polarity of the word combination as a whole, is assigned. This polarity degree is an integer from the set {-3, -2, -1, 1, 2, 3}. For example, the polarity degree 2 in the second column of the Table 6 can be interpreted as strong positive polarity (SP) or weak positive polarity modified by intensity (WP + I). This intensity is introduced into the given combination by another word, which can precede or follow the word with weak positive polarity. Table 6 illustrates examples of most often used word combinations for *K* from 2 to 4 together with their interpretation and resulting polarity degree.

Table 6. Polarity degree determination of words combinations with various code lengths (SP+I is Strong Positive + Intensity, SP or WP+I represents Strong Positive or Weak Positive + Intensity and WP is Weak Positive. Similarly, it holds for negative polarity.)

Interpretation	SP + I	SP or WP + I	WP	WN	SN or WN + I	SN + I
K = 2	48	80, 41	10, 32, 23	20, 31, 13	90, 42	49
K = 3	480,408	800, 410, 401	100, 320, 230, 302, 203	200, 310, 130, 301, 103	900, 420, 402	490, 409
K = 4	4800, 4080, 4008	8000, 4100, 4010, 4001	1000, 3200,2300, 3020,2030, 3002,2003	2000, 3100,1300, 3010,1030, 3001,1003	9000, 4200, 4020, 4002	4900, 4090, 4009
polarity	**3**	**2**	**1**	**-1**	**-2**	**-3**

According to the second column of the Table 6, the polarity degree 2 (with its interpretation SP or WP + I) for *K=4* represents two basic alternatives. The first possible alternative is represented by a strong positive word (8), which is complemented by neutral words (8000). The second possibility is a weak positive word (1) followed (within the same combination) by word increasing polarity intensity (4) and they are complemented by two neutral words in order to form a combination of the given length (4100). These words having non-zero code can be differently ordered within the given word combination (e.g. 4010, 4001).

Table 6 is not presented in its entirety. It only illustrates the most often employed combinations. For example, the second column can be completed with other combinations, for example a weak positive word can be followed by a word increasing polarity intensity (1400, 1040 and 1004).

5 Implementation of the Opinion Classification Method

The presented design of the method of opinion classification has been implemented as well. The implementation within OCS (Opinion Classification System) was used to experiment with the designed method. The OCS is a server application with two interfaces – one interface for “guest” users and another one for “admin” users. Expected competencies of the guest users are: initialization of opinion classification of a selected text and changing the value of the dynamic coefficient, if it is necessary. The admin user has the same competencies as guest but he/she can also create and edit the classification vocabulary. When the OCS system detects a new word within the processed text, it offers to admin the possibility to insert this new word into the classification vocabulary. The admin can decide whether to insert this unknown word (the word has subjectivity) into the vocabulary or not (the word has no subjectivity). This implementation has been realized in the programming language PHP and it is available on the URL http://mk51.wz.cz/. More information about this implementation can be found in [7].

The implementation was tested on the set of discussion contributions from the portal http://www.mobilmania.sk. This portal focuses on mobile telephones evaluation. Our tests were focused on the discussion thread related to reviews of the mobile telephone LGKU990. The set of contributions used for testing purposes contained 1558 words and 236 lexical units (combinations). The structure of the classification vocabulary was the following: 27 positive words, 27 negative words, 10 negations and 11 words, which increased the intensity of polarity. The evaluation was based on the comparison of results achieved by the OCS system and results obtained from an expert. The expert provided logical analysis of contributions taking into account the structure and meaning of particular sentences. The resulting precision of the implementation OCS according to introduced tests was 78,2%, which is arithmetical average of precision of OCS on positive contributions (86,2%) and on negative contributions (69,2%), what can be seen in Table 7.

Table 7. Results of experiments with the implementation OCS

	OCS result	Expert result	Precision
positive	29	25	0,862
negative	26	18	0,692

We can see in the table, that the OCS implementation classified some neutral or even negative (positive) contribution to the positive (negative) opinion category. There are 4 mistakes in the classification of 29 contributions as positive opinions. For example, the sentence ‘Also my old Sony Ericsson makes *better* photos’ was classified to positive opinion category because of the positive word ‘better’ and lack of ability of OCS to identify hidden irony of this sentence.

The opinion classification is sometimes very complicated not only due to the irony. Complications can arise from indirectly expressed opinion as well. For example, let us consider the sentence ‘I would not buy other brand’. It contains only neutral words and negation without positive or negative word, which this negation is related to. Therefore, the OCS classified this sentence to the neutral opinion class.

6 Conclusions

The automatic opinion classification definitely belongs to up-to-day research agenda. There is a great potential of using the opinion classification within web discussion portals as a service not only for ordinary users (consumers) but for business-entities or organizations (Internet marketing) as well. The application of opinion classification can offer help supporting common users in decision making. Similarly, it can offer some services to business-entities and organizations (e.g. political parties, subjects of civil services, printed and electronic media, marketing agencies, etc.), for example the prediction of the development of society feelings or measuring degree of freedom of media. From this point of view, it is very promising research field.

The achieved precision of our implementation (78,2%) can be perceived as a relatively good result considering the beginning stage of development. During next research stage, this implementation should be improved in order to perform deeper analysis of the given text and to provide more precise opinion classification. Higher precision can be achieved by means of irony and ambiguity detection. Also, it would be appropriate to test the improved implementation within the more extensive testing environment setting.

Acknowledgements. This work was supported by the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic within project 1/0042/10 “Methods for identification, annotation, search, access and composition of services using semantic metadata in support of selected process types”(50%). This work is also the result of the project implementation Development of the Centre of Information and Communication Technologies for Knowledge Systems (project number: 26220120030) supported by the Research & Development Operational Program funded by the ERDF (50%).

References

1. Barla, M., Bieliková, M.: On Deriving Tagsonomies: Keyword Relations Coming from Crowd. In: Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems, LNAI, vol. 5796, Proc. of the ICCCI 2009, Poland, 309–320, Springer, Heidelberg (2009)
2. Catena, A., Alexandrov, M., Ponomareva, N.: Opinion Analysis of Publications on Economics with a Limited Vocabulary of Sentiments. International Journal on Social Media. MMM: Monitoring, Measurement, and Mining, Vol. 1, No. 1, 20-31 (2010)
3. Ding, X., Liu, B., YuA, P. Holistic Lexicon-Based Approach to Opinion Mining. Proc. of the Int. Conf. on Web Search and Web Data Mining WSDM'2008, New York, NY, USA, 231-240 (2008)
4. Heer, J., Boyd, D.: Vizster: Visualizing Online Social Networks. Proceedings of the IEEE Symposium on Information Visualization INFOVIS'2005, Washington, USA, 5-13 (2005)
5. Kaurova, O., Alexandrov, M., Ponomareva, N.: The Study of Sentiment Word Granularity for Opinion Analysis (a Comparison with Maite Taboada Works). International Journal on Social Media. MMM: Monitoring, Measurement, and Mining, Vol. 1, No. 1, 45-57 (2010)
6. Lukáč, G., Butka, P., Mach, M.: Semantically-enhanced extension of the discussion analysis algorithm in SAKE. In: SAMI 2008, 6th International Symposium on Applied Machine Intelligence and Informatics : January, 2008, Herľany, Slovakia, 241-246 (2008)
7. Machová, K., Krajč, M.: Opinion Classification in Threaded Discussions on the Web. Proc. of the 10th annual international conference Znalosti 2011, Stará Lesná, Publisher: FEI Technická univerzita Ostrava, Czech Republic, 136-147 (2011)
8. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundation and Trends in Information Retrieval, Vol.2, No.1-2, 1-135 (2008)