Ersch. in: International Journal of Data Mining, Modelling and Management; 5 (2013), 2. - S. 138-157
http://dx.doi.org/10.1504/IJDMMM.2013.053693

What do you think about this photo? A novel approach to opinion and sentiment analysis of photo comments

Slava Kisilevich* and Christian Rohrdantz

Department of Computer and Information Science,
University of Konstanz,
P.O. Box 78, 78457 Konstanz, Germany
E-mail: slaks@dbvis.inf.uni-konstanz.de
E-mail: christian rohrdantz@uni-konstanz.de
*Corresponding author

Veronica Maidel

School of Information Studies,
Syracuse University,
343 Hinds Hall, Syracuse,
NY 13244-4100, USA
E-mail: vmaidel@syr.edu

Daniel Keim

Department of Computer and Information Science,
University of Konstanz,
P.O. Box 78, 78457 Konstanz, Germany
E-mail: keim@dbvis.inf.uni-konstanz.de

Abstract: We propose a practical unsupervised approach to opinion and sentiment analysis of photo comments with a real-valued strength orientation. We extract two types of opinions: opinions that relate to the photo quality and general sentiments targeted towards objects depicted on the photo. Our approach combines linguistic features for part of speech tagging, traditional statistical methods for modelling word importance in the photo comment corpus (in a real-valued scale), and a predefined lexicon for detecting negative and positive opinion orientation. In addition, we apply a semi-automatic photo feature detection method and introduce a set of syntactic patterns to resolve opinion references. The results of our user study among 49 non-expert participants of different ages showed no statistical differences between user evaluation and the algorithm.

Keywords: photo comments; opinion and sentiment analysis; inter-rater agreement; intraclass correlation coefficient; data mining.

Reference to this paper should be made as follows: Kisilevich, S., Rohrdantz, C., Maidel, V. and Keim, D. (2013) 'What do you think about this photo? A novel approach to opinion and sentiment analysis of photo comments', *Int. J. Data Mining, Modelling and Management*, Vol. 5, No. 2, pp.138–157.

Biographical notes: Slava Kisilevich received his MSc in Information Systems Engineering at Ben-Gurion University of the Negev. He currently pursues his PhD in the Information Visualisation and Data Analysis Research Group of Professor Daniel Keim at the University of Konstanz. Christian Rohrdantz received his MSc in Information Engineering at the University of Konstanz, Germany. He currently pursues his PhD in the Information Visualisation and Data Analysis Research Group of Professor Daniel Keim at the University of Konstanz.

Veronica Maidel received her MSc in Information Systems Engineering at Ben-Gurion University of the Negev. She currently pursues her PhD in Information Science and Technology at Syracuse University.

Daniel Keim is Full Professor and Head of the Information Visualisation and Data Analysis Research Group at the University of Konstanz, Germany. He has been actively involved in information visualisation and data analysis research for more than 15 years and developed a number of novel visual analysis techniques for very large datasets with applications to a wide range of application areas including financial analysis, network analysis, geo-spatial analysis, as well as text and multimedia analysis.

This paper is a revised and expanded version of a paper entitled 'Beautiful picture of an ugly place. Exploring photo collections using opinion and sentiment analysis of user comments' presented at 2010 International Multiconference on Computer Science and Information Technology (IMCSIT), 18–20 October, Wis³a.

1 Introduction

With the fast development of user-centred internet technologies, we witness a rapid growth of web resources, which not only allow users to obtain, but also to generate their own textual information. This leads to dramatic improvements of products and services. For example, nowadays it is difficult to imagine that we would book a hotel room without checking the hotel's overall ranking or without reading comments previously written by other users. We are also less inclined to buy a product without reading comments or ratings about its quality. In fact, written opinions have become essential components in decision-making processes and are common in almost all parts of our life. They are essential parts of blogs, news, financial market reports, product reviews, etc. However, textual information generated on the web grows almost at an uncontrollable pace, and manual skimming through user opinions has become a time consuming process.

A typical task in opinion mining is to determine whether a document (review, comment) is bearing a positive or negative connotation (Hatzivassiloglou and McKeown, 1997; Turney, 2002; Dave et al., 2003; Salvetti et al., 2004; Das and Chen, 2007; Fahrni and Klenner, 2008; Argamon et al., 2007). If either connotation is present, the task can be formulated as a classification problem with two class labels (positive and negative) (Liu, 2009). Three different kinds of approaches have been used: unsupervised (Turney, 2002), semi-supervised (Argamon et al., 2007) and supervised ones (Gamon, 2004;

Salvetti et al., 2004; Chesley et al., 2006; Das and Chen, 2007; Drake et al., 2008; O'Hare et al., 2009). Supervised machine learning approaches perform well if sufficient labelled training data exist (for example, in the domain of movie reviews, users assign ranks to movies along with their written opinions). However, in domains where labels are not easily acquired or where opinion orientation is measured on a real-valued scale (Subrahmanian and Reforgiato, 2008), unsupervised approaches are more favourable.

In this paper, we consider the problem of opinion and sentiment analysis of users' comments written for photos that are uploaded to photo sharing websites. Detailed inspection of user comments revealed that comments are noisy, relatively short, and contain only few negations. They may be written in any language, contain arbitrary syntactic structures and typos. Moreover, they may contain a mixture of opinions on the quality of the photo (usually positive) 'great shot', 'nice picture' and sentiments or moods expressed towards objects depicted on the photo ('sad place'). Further observations revealed that written opinions are mostly accompanied by adjectives, which is in accordance with past findings (Wiebe et al., 1999; Wiebe, 2000). As mentioned above, a widely used approach is to classify documents using a binary classification. This approach seems inappropriate in our case for two reasons:

1 Photo comments have two subjects of opinions (opinions on the photo and sentiments towards objects). Consequently, we will lose valuable information if the overall score will be a mixture of two opinion scores.

2 Since most of the opinions are positive, we will end up with most of the photo comments classified as positive.

In order to provide a workable essential-feature analysis, we propose two improvements over existing approaches. We extract two types of opinions:

1 Opinions that relate to the photo quality.

2 General sentiments (GSs) targeted towards objects depicted on the photo.

Supervised machine learning approaches are not feasible in our case since it is very hard to find agreements between human annotators on a real-valued scale, e.g., the difference in opinion strength between 'great shot' and 'amazing photo' cannot be clearly defined. For that reason, we propose an unsupervised approach for opinion scoring using concepts of word importance based on statistical properties derived from the field of information retrieval (Salton and Buckley, 1988) and using concepts of Zipf's speech regularity (Zipf, 1949) and semantic differentiation (Osgood, 1957).

Based on the observations described above, we generated our own lexicon of adjectives extracted from the corpus of user comments, and analysed its usage with respect to photo quality opinions and GSs, as well as their usage by commenters. We found that in the majority of cases, adjectives are used directly with the subject of the opinion ('great shot') and that the most frequently used adjectives are the same, even if different regions of the world are considered with photos of different subject matters. The latter suggested that a finite lexicon of adjectives could be used for opinion and sentiment analysis of photo comments in many regions.

The main contributions of the paper can be summarised as follows:

1 Our model is based on the corpus extracted from users' photo comments.

2 We construct and employ a finite lexicon of opinion words in contrast to the majority of approaches in which seed lists are used to infer scores of unknown opinion words. Therefore, we complement a predefined opinion lexicon with the mentioned adjective weighting model.

3 We develop a model that consists of two types of scores: *opinion* regarding the photo and *sentiment* towards the subject of the photo. For this purpose, we suggest a semi-automatic extraction of photo features and a set of syntactic opinion reference patterns.

4 We model the orientation strength based on word distributions without using any external dictionaries, while the semantic orientation (positive or negative) of a word is determined by the predefined lexicon of positive and negative opinion-bearing words.

5 We provide a continuous scale for opinion and sentiment orientation.

6 With our approach, we allow for dynamic updates of scores when new comments are added to the system, which makes the whole method readily applicable in real-world tasks (Kisilevich et al., 2010).

7 As the basis for our approach, we conducted a carefully designed extensive user study. Apart from demonstrating the performance of our approach, the user study provided further interesting insights on how users perceive opinions and sentiments in photo comments.

2 Related work

Existing approaches in the context of opinion analysis can be broadly divided into several categories. The following categories are closely related to our work: *opinion classification*, *lexicon generation*, and *feature-based opinion analysis*. A more detailed overview can be found in Liu (2009).

A Naïve Bayes classifier was used in Salvetti et al. (2004) for classifying movie reviews, while Das and Chen (2007) used Naïve Bayes as one of five classifiers with majority voting. A support vector machine (SVM) classifier was used by Gamon (2004) for classifying customer feedback data. O'Hare et al. (2009) applied SVM on financial blogs. An unsupervised approach for review classification was applied in Turney (2002) based on the calculation of pointwise mutual information (PMI) among potential opinion phrases in a large-scale web corpus. Subrahmanian and Reforgiato (2008) proposed a real-valued scale opinion orientation based on a classification of adverbs, different verb categories and complex relationships of adverbs, adjectives and verbs in the text. The mentioned classifications are used to separate comments according to their opinion orientation or in order to asses opinion strength. In our approach, we additionally separate opinions about the photo quality from sentiments about the content.

Additional approaches to learn the semantic orientation of words utilise external resources like WordNet (Fellbaum, 1998) by measuring relative distance of an arbitrary word to words 'good' and 'bad' (Kamps et al., 2004) or by utilising a random walk model on the graph of word relations (Hassan and Radev, 2010). Esuli and Sebastiani

(2006) generated a dictionary called SentiWordNet using WordNet with three sentiment scores (positive, negative and objective) for each WordNet synset. Other approaches rely on seed lists containing words with a known semantic orientation and search corpora for specified adjective-adjective relations (Hatzivassiloglou and McKeown, 1997), adjective-product feature relations (Qiu et al., 2009; Jijkoun et al., 2010) or bag-of-word vector space similarities (Sahlgren et al., 2007). Chesley et al. (2006) used a Wikipedia dictionary to determine the polarity of adjectives. We use a predefined opinion lexicon, the internet general inquirer lexicon (http://www.webuse.umd.edu:9090/), and complement it with a statistically motivated adjective weighting model.

In addition to the approaches that try to detect the sentiment of sentences or even documents as a whole, the task of feature-based analysis is to investigate to which feature (e.g., entity, topic, attribute) sentiments or opinions refer. Some of the feature-based analysis methods use distance-based heuristics (Ding et al., 2008; Oelke et al., 2009). The closer an opinion word is to a feature word, the higher its influence on the feature is. Other approaches exploit advanced natural language processing methods, like dependency parsers, to resolve linguistic references from opinion words to features. Popescu and Etzioni (2005) extract pairs (opinion word, feature) based on ten extraction rules that work on dependency relations involving subjects, predicates and objects. Riloff and Wiebe (2003) use lexico-syntactic patterns in a bootstrapping approach to resolve relations between opinion holders and verbs for subjectivity classification. In our approach, photo features are extracted from the comments based on word distribution characteristics across regions. Then, opinions referencing these photo features are identified according to predefined part-of-speech patterns.

3 Development of photo comments corpus

In this section, we outline the photo comment collection, the creation of the corpus, and the preprocessing techniques.

3.1 Data collection

We collected photo comments from Flickr (http://www.flickr.com/), the largest web community for photo and web sharing, using its publicly available API. Between June 2009 (as part of another project) and the end of April 2010, we collected metadata for about 90 million geotagged photos from about 7.6 million users.

3.1.1 Region selection

Five regions (Dachau, Auschwitz, Wisła, Krakow and Warsaw) were defined for analysis. The rationale behind selecting these regions pertains to the following three goals:

1 To find differences in comment types between regions.

2 To find differences in the usage of parts of speech (adjectives and nouns).

3 To build a model that represents the nature of photo comments.

We assumed that Dachau and Auschwitz concentration camps should contain special kinds of comments (negative emotions) that would differ from comments in general

tourist locations. Wisła, we assumed, is a neutral region without many attractions while Krakow and Warsaw were selected as large Polish cities that include many tourist attractions. Table 1 summarises the statistics related to the selected regions.

Table 1 Statistical information related to five regions selected for analysis

Region	*Area*	*# Commented photos*	*# Owners*	*# Commenters*	*# Commented photos after preprocessing*
Krakow	120 km^2	8,127	1,257	23,045	4,214
Warsaw	60 km^2	8,690	1,140	22,695	4,098
Wisła	43 km^2	117	39	603	56
Auschwitz	12 km^2	505	138	1,687	311
Dachau	14 km^2	329	121	1,062	179

3.1.2 Preprocessing

Having manually examined hundreds of user comments, we found a similarity to blogs (Chesley et al., 2006), where opinions are stated in the beginning of the paragraph. Similar to blogs, the same user can write several comments about the same photo, but usually the first comment contains the opinions and sentiments, while subsequent comments mostly include neutral information like responses to comments of others or the photo owner. The following example shows two comments from the same user. In the first comment, there is an expression of sentiment ('powerful place and story'). The second comment was made after the owner of the photo wrote his response.

1 *This is great. I visited Dachau, but don't remember this part. But I hear they have added some things in the last 5 years. Powerful place and story, thanks for sharing.*

2 *I was there about 8 years ago and I don t recall this hall way. Was this one of the houses, or near the main complex where the museum and films were?*

As already mentioned, the owner of the photo can also participate in the discussion about his own photo. The following is a short example of two comments written by the owner of the photo to people as a response to their comments.

1 *Thanks for the comments. I also found the colors both beautiful and chilling...a very creepy place for sure.*

2 *Thanks! I was fortunate to actually capture the impression it made on me standing there in person.*

In this case, his opinions can introduce a certain bias, which suggests that comments of the photo owner should be excluded from the analysis.

For every region, we selected photos that contain at least one comment. We removed HTML tags and irrelevant sections (URL links, invitations to join a group). Next, we applied a language guesser to remove comments written in languages other than English and applied Stanford POS tagger (Toutanova and Manning, 2000) on the remained comments. Table 1 shows the number of remaining commented photos after the preprocessing.

4 Method

4.1 Definitions

Different terminology definitions are provided in the sentiment and opinion analysis literature. The terminology used in this paper mostly follows the definitions given in Liu (2009), but makes a clear distinction between opinions and sentiments. The important terms and their definition for this paper:

- Photo feature: Nouns that describe the photo features – attributes, components or characteristics of the photo, e.g., 'shot', 'photo', 'colour', 'composition', 'light'. Photo features in our case are usually related directly to the quality of the photo. It is common to distinguish between explicit and implicit features, i.e., features that are mentioned in a sentence and features that are not explicitly mentioned but implicitly referenced.
- Orientation: The semantic orientation of a word or a comment as a binary categorical variable with the parameter values 'negative' and 'positive'. Sentences or words that cannot be assigned to one of these two categories are implicitly rated as 'neutral' and ignored in the further analysis.
- Orientation strength: The numerical strength of the orientation value ranging from 0 to ∞ in absolute numbers, whereas negative orientations are indicated by the algebraic sign '-'.
- Photo opinions (PO): Negative or positive user statements, that clearly refer to photo features of a certain photo, are summarised as the respective PO. They express the users' opinions on the technical and artistic photo quality. For simplicity, we will only speak of *opinions* when we refer to *POs*.
- General sentiments (GS): Negatively or positively connoted user statements that cannot be attributed to a photo feature. As implied by the denotation, the GS shall capture orientation statements that have a broader nature than opinions, i.e., sentiments and emotions that are evoked by the photo content. For simplicity's sake, we will only speak of *sentiments* when we refer to *GSs*.

4.2 Corpus-based lexicon generation

Opinion mining is heavily dependent on an opinion lexicon. There are two common approaches to generating a lexicon, the dictionary-based and the corpus-based approach. The former is based on bootstrapping a seed of opinion words from dictionaries like WordNet (Fellbaum, 1998), SentiWordNet (Esuli and Sebastiani, 2006) or Wikipedia (http://www.wikipedia.org/), the latter is based on the corpus and, thus, inherently domain dependent. We extend an existing general lexicon, the Internet General Inquirer lexicon, and adapt it to our task using an adjective-weighting model.

We applied a corpus-based lexicon generation due to different reasons:

1 We want to generate a new lexicon in the domain of photo comments since currently, at least to our knowledge, no such lexicon is publicly available.
2 Dictionaries may supply only a binary opinion orientation, while our task is to model opinion orientations on a real-valued scale.
3 We want to investigate statistical properties of words used for commenting.

In order to acquire word distributions, we extracted adjectives and nouns from the corpus, counted their occurrences in the five selected regions separately, and sorted them according to their frequency from the highest to the lowest. Nouns were extracted in order to learn what words are commonly used as photo features. We used the Yago-Naga stemmer (http://www mpi-inf mpg.de/yago-naga/) to convert all nouns into a singular form.

To minimise the bias of some very active commenters, we counted word occurrence only once for each person for each region. The reason why we selected five separate regions is because word occurrences may differ due to different subject matters. Moreover, the number of commented photos is different from region to region and the word distribution would inevitably be biased towards words used in regions with many comments.

An inspection of the adjective distribution is quite surprising: The words *great, nice* and *beautiful* are the most frequent and equally ranked adjectives in all five regions[1]. Among the 100 frequent adjectives, 36 adjectives are unique, 58% of the adjectives are found in more than one region and 42% of these frequent adjectives are found only in one region. This suggests that the vocabulary that people use to express opinions or sentiments is relatively small and contains many common words even if the context of the photos is very different (e.g., Dachau concentration camp and Nature).

Next, we obtained the slope coefficients of word frequencies to check for existence of Zipfian distribution. The slope coefficients are the following: Krakow (–1.138), Warsaw (–1.136), Auschwitz (–0.988) and Dachau: (–0.95) (Wisła was excluded because it does not have enough words for a reliable slope estimation). The results show that Zipf's law holds true not only for the English language as a whole but also for a particular parts of speech usage in photo comments.

4.3 The adjective weighting model

Having shown the statistical properties of the distributions of adjectives in the photo comments corpus, we are now ready to discuss the linguistic interpretation of adjective usage and propose an adjective weighting model for opinion orientation.

It was shown in past research that there is a strong correlation between the presence of adjectives and opinions (Wiebe et al., 1999; Wiebe, 2000). Indeed, a careful analysis of photo comments showed that people often use short sentences like 'great photo', 'nice picture', 'sad place' to express their opinions or sentiments. The analysis also showed that the number of positive adjectives used in photo comments is higher than the number of negative adjectives and that overall, the number of positive comments is much higher than the number of negative comments. Any lexicon of positive and negative words will show that the words 'great' and 'nice' are positive. However, it is difficult to estimate which of these two words is 'more positive than the other' using lexical features alone. Osgood (1957) pointed out that a difference in 'feeling-tone' exists even between synonyms such as 'good' and 'nice', but people are unable to verbalise the difference.

One of the simple approaches is to treat all positive words as equally positive, assigning a score of 1 for every occurrence of a positive word and counting the total number of positive words in a sentence or a document. Likewise, the negative words could be assigned a score of –1. Consequently, the final orientation of a sentence or a document would be the overall score (positive or negative) calculated by addition of

all positive and negative scores (Turney, 2002). This approach was used in previous research in the context of classifying the documents into positive or negative classes. Our task is different since we are interested in not only classifying the documents but also in ranking them according to the opinion or sentiment strength. As we mentioned, the majority of comments are relatively short and according to the statistics acquired from the five regions that we investigated, the vocabulary that people use to express their opinions or sentiments is relatively small. Thus, we hypothesise that a mere counting of positively and negatively oriented words will result in lack of sensitivity between the ranked comments. We claim that opinion or sentiment words should be scaled on a continuous scale denoting the difference in opinion or sentiment strength between those words. Therefore, we base our claim reflecting upon the seminal work of Osgood 'The measurement of meaning' and using the least effort principle as well as word distribution regularity presented by Zipf in his 'Human behaviour and the principle of least effort'. As we showed in Section 4.2, the orderliness of word distribution is preserved not only for particular parts of speech but also in every region. This indicates that even in special cases where photo comments are written by non-native speakers of English as well as by native English speakers, the fundamental principle that governs the word usage in a language is preserved even if a person is not aware of its existence as suggested by Zipf (1949). Moreover, if any regularity or law did not govern the word usage it could mean that people do not attach any meaning to what they are saying or that they do not differentiate between words that describe the same concept. In the former case, we could observe a completely random word occurrence, in the second case we could observe that the frequency of word usage is the same no matter what word is used.

Similar to Osgood's measurement of meaning to compare 'the output of two different subjects' measuring similarity or difference in meanings of a term, our goal is to quantitatively measure the opinion or sentiment strength. Unlike Osgood that builds differential scales for every concept (good-bad, slow-fast), we utilise Zipf's fundamental law of word usage by comparing how words that denote the same concept (positive or negative in our case) are used by people. This can be compared to a TF-IDF measure often used in information retrieval (Jones, 1972; Salton and Buckley, 1988). Let us assume that 'good' and 'nice' are the two words with equal frequency (TF is equal). According to TF-IDF, the least important word is the one which is found in most of the documents. Similarly, the word with the highest importance is the word that is found in the least number of documents. In this case, one of the words, let us say 'good', which is found in most of the documents will receive a lower score than the word 'nice'. Similarly to TF-IDF, in our case, the most frequent word is the one which is found in most of the comments. Thus, its score will be lower than the score of the next most frequent word.

We define the word opinion strength w_{oo} using the principles of word importance as defined in the TF-IDF measure and word distribution properties of Zipf's law as follows:

$$w_{oo} = orientation(w) * lg\left(\frac{f_{w,r=1}}{f_w} + 1\right) \quad (1)$$

where $orientation(w)$ is a function which assigns 1 if the word w is positive and –1 if it is negative, $f_{w,r=1}$ is the frequency of the word having the rank 1 (a most frequent word) and f_w is the frequency of the word w in the whole corpus.

The difference between TF-IDF and our approach is that the importance of the word in TF-IDF is measured for every word independently, while the opinion orientation score is calculated relatively to the most frequent word in the corpus. Thus, if the most frequent word is 'great' with a frequency of 1,469 and the word ranked second is 'nice' with a frequency of 864, 'great' will receive a score of 0.30 (lg (1,469/1,469 + 1)), while the score of 'nice' will be 0.43 (lg (1,469/864 + 1)). One is added to *log* to avoid a zero score of the most frequent word.

We should note, that the word frequency in equation (1) is absolute and can be applied to five regions separately. In order to create a global model that takes into account different word distributions, we need to find the relative order of all words from five regions. We proceeded as follows:

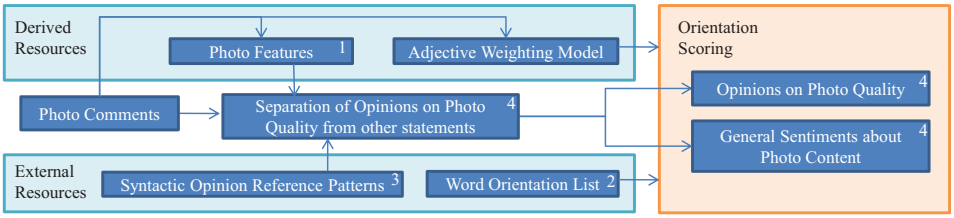
- We calculated a ratio $\frac{f_{w,r=1}}{f_w}$ for every word.
- An average of ratios for every word was calculated taking these ratios for the same word $w_{i,n}$ from every region n.
- If the word $w_{i,n}$ was not found among the lexicon of the region n, its ratio was assumed to have the ratio of the last word in the lexicon of the region n.

After building a weighted ratio for every word, we applied equation (1) to obtain the global adjective weighting model.

4.4 *Automatic opinion and sentiment analysis*

The automatic opinion and sentiment analysis consists of several interdependent steps as outlined in Figure 1. The analysis relies on both resources derived from the photo comment corpus itself and external resources. The details are provided in the following subsections.

Figure 1 Interdependence of the different core text analysis processes (see online version for colours)



Note: The numbers correspond to the paragraphs in Section 4.4, where details are provided.

4.4.1 *Photo features*

In order to determine which opinions relate to the photo, first a list of photo features had to be compiled. For this purpose a term extraction method was created that exploits certain characteristics of photo features:

1 Features usually correspond to nouns.

2 Features should not depend significantly on the photo location.

3 Features should be frequent in photo comments.

Consequently, all nouns were extracted, that appeared in photo comments of at least four out of five locations and finally the 100 most frequent among these terms were extracted as candidate photo features. The list was then manually revised and finally, 50 out of these nouns were considered in the analysis as photo features. The top ten frequent nouns present in at least four locations were, in decreasing frequency order, 'shot', 'photo', 'colour', 'composition', 'light', 'picture', 'capture', 'love', 'image', 'work'. Here, 'love' is one example that was manually deleted. In this case we could observe that the high frequency of the noun 'love' was due to a repeated error of the part-of-speech tagger, when occurrences of the verb 'love' in very short sentences (e.g., 'love it!') were misclassified as nouns.

Implicit features: A number of very short sentences implicitly refer to the photo quality without explicitly mentioning a photo feature (e.g., 'i love it.', 'well done.', 'very nice.'). The common characteristic of such sentences is that they are very short and do not contain any nouns, i.e., do not contain any explicit target word for sentiments or opinions. Therefore, for very short sentences (less than six words) that did not contain any nouns, it was assumed that they implicitly related to the photo quality.

4.4.2 The word orientation list

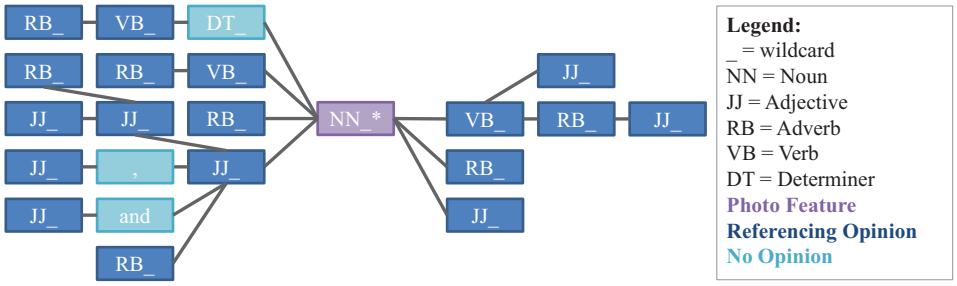
A manually enhanced version of the widely used internet general inquirer lexicon was used as a word orientation list. It was applied to determine the orientation of the word and incorporate it into equation (1), i.e., +1 for positive, –1 for negative and 0 for neutral words (not contained in the orientation list). All the words which were not contained in the adjective weighting model (Section 4.3), were allocated the weight of 1, because they either had not appeared in the photo comments or because they belonged to a different part-of-speech category.

4.4.3 Syntactic opinion reference patterns

In order to detect references of opinion words to photo features, a set of syntactic opinion reference patterns was defined, based on linear word order part-of-speech sequences[2]. A very simple example is the pattern 'JJ NN', which stands for an adjective (JJ) directly followed by a noun (NN). In this case, we could be sure that the adjective referred to the noun. Hence, if the noun is a photo feature then the adjective and its orientation can be assigned to this feature. While in theory recursive patterns of arbitrary length (e.g., JJ* NN) are possible in natural language, in practice such patterns do not appear to a noteworthy extent in the domain under investigation. When we defined the pattern set, we started with including some very obvious cases like 'JJ NN' and 'NN VB JJ' and then skimmed through the data in search for further patterns. We could observe that the limited pattern set we defined covered the vast majority of cases. To verify the observation, we randomly drew sentences from the corpus until having encountered 100 opinion reference examples. While 90 were correctly covered by our patterns, no false positives occurred. The whole pattern set is provided in Figure 2. One main advantage is that the patterns encode the available linguistic knowledge about opinion

references without requiring the time-consuming parsing of a full syntax structure tree or a typed dependencies graph.

Figure 2 Syntactic opinion reference patterns (see online version for colours)



Note: Word order patterns go from left (before photo features) to right (after photo features), the distance to the photo feature indicates the exact position.

Our syntactic reference patterns cover most of the cases that other approaches detect with dependency parses. This is because in English, adjectives are usually very close to the nouns they refer to or modify. Only very exceptional and infrequent cases like a relational phrase, e.g., the hypothetical sentence "the photo, that shows a tree, is really nice" cannot be resolved by our means. In case of verbs, our approach is not able to distinguish explicitly whether the feature is the subject or the object of the verb. In our tests, however, we could observe that this is not a problem. In addition, our method is less error-prone than dependency parsing, especially when applied to less formalised and sometimes sloppy and incorrect writing, as in user-generated content.

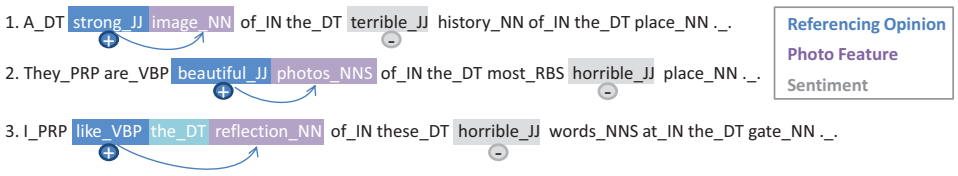
4.4.4 Identification and separation of POs and GSs

A crucial part of the automatic text analysis is the detection and separation of (1) opinions about the photo quality (PO) and (2) general sentiments expressed about the photo content (GS).

The first part (1) is based on the extraction of photo features and the mapping of opinion statements to photo features. The described set of syntactic opinion reference patterns was applied for this mapping. For each photo feature in a sentence, all words were extracted that describe the feature according to one of the syntactic opinion reference patterns. The orientation scores of these words were then summed up to yield a PO value. In this process, a simple heuristic is used to invert the orientation of negated words.

Accordingly, part (2) is based on all sentiment expressions that could not be attributed to photo features during Step (1). This means that all the words that do not refer to photo features were considered and their orientation scores were summed up to yield a GS value. Figure 3 provides some example sentences extracted from the photo comments, which contain both POs and GSs. In all three cases, we have positive opinions ('strong', 'beautiful', 'to like') on the artistic quality of the photo, represented by the photo features ('image', 'photos', 'reflection'). The corresponding part-of-speech sequences are included in the syntactic reference patterns. The remaining opinion words ('terrible', 'horrible') that could not be attributed to a photo feature are consequently considered as referring to GSs on the photo content.

Figure 3 Three POS-annotated example sentences extracted from the photo comments (see online version for colours)



Note: Each of these sentences contains both a PO and a GS.

It should be noted that clauses related to GSs are falsely classified as POs only in very rare cases. The opposite situation, where POs are falsely classified as GSs, could be observed in a couple of cases, due to different reasons (missing photo feature, implicitness).

5 Experimental evaluation

The goal of the experimental evaluation is to compare the performance of the proposed approach to the performance of human evaluators and to determine the factors that influence non-expert evaluators during opinion and sentiment strength assessment.

5.1 Design

A total of 78 participants were recruited to participate in the user study through Amazon's mechanical turk (http://www mturk.com/) of which 49 participants (31 females, 18 males) completed the assignment. The age of participants ranged from 18 to 67 (mean 31.3, std. 11.14, median 28). The user study lasted for one week and was restricted to users from the USA. Each person that accepted the assignment, received a questionnaire and a set of five text files containing user comments, which were gathered from photos randomly selected by the automatic procedure. The evaluator had to judge comments according to the criteria (opinion or sentiment), manually assigned to him/her by the user study manager. The evaluation procedure consisted of three steps.

In the first step, the participants provided some demographic information about themselves, such as age and gender. We also asked the non-native English participants to assess their level of English (basic, intermediary, high). 37 people were native English speakers. Six people stated that their level of English was high, five stated that their English was on an intermediary level, and one person stated to have basic knowledge of English.

In the second step, the participants had to read the comments in the files and rank them according to the opinion or sentiment strength from the most positive opinion or sentiment to the most negative. In total, 60 sets of comments were prepared for five regions: Auschwitz, Dachau, Krakow, Warsaw, and Berlin. Berlin was chosen as an additional region on which we applied our global weighting model. Sentiments assessment criteria was applied on comments from Auschwitz, Dachau, Berlin while opinions assessment criteria was applied on comments from Krakow, Warsaw, and Berlin. Every set contained comments from five photos. Every set was generated by

randomly selecting photos from a particular region. 49 participants evaluated 137 sets reading 685 photo comments, which yields 2.79 sets per participant on average and 2.28 evaluations per set (some sets were evaluated by three participants).

The third step included eight closed-ended questions (see Table 3) to assess the additional factors that might have influenced the evaluator and one open-ended question to be filled by the evaluator in case there was a factor that was not mentioned. A five point Likert scale was used for the closed-ended questions ranging from *strong disagree* to *strong agree*.

5.2 Method

Kendall's tau rank correlation was used to assess the degree of inter-rater agreement (IRA) between the ranks produced by the algorithm and the ranking of users. We applied the intra-class correlation coefficient (ICC) (Shrout and Fleiss, 1979) to assess the differences in the opinion or sentiment scores assigned by the algorithm and the users. Since the users were asked to provide only ranks and not scores, an item ranked at the *i*th place by the user, got the score assigned to it by the algorithm. This allowed us to avoid unnecessary complications with differences in user scoring. In all cases, the average IRA and ICC was calculated for every set for all users and then, the averaged IRA and ICC were averaged across sets in every region, across sets belonging to the same evaluation criterion (all sentiments, all opinions), and across all sets (All) without regard to a criterion (see Table 2). Finally, the Mann-Whitney U two-tailed test with significance level α of 0.05 was used to answer the question whether the rankings and the score differences between the algorithm and human evaluators are not statistically significant, i.e., whether the performance of the algorithm and the performance of the human evaluators are the same (the null hypothesis). The answers to the closed-ended questions were numerically encoded from –2 (*strong disagree*) to 2 (*strong agree*), and the mean, standard deviation, and median were calculated (see Table 3).

Table 2 Algorithm-user and user-user inter-rater agreement (IRA) and ICC

Dataset	*Test*	*IRA (Avg./SD)*	*ICC (Avg./SD)*
Auschwitz (sentiments)	Alg-User	0.325 ± 0.419	0.457 ± 0.436
	User-User	0.164 ± 0.436	0.230 ± 0.481
Dachau (sentiments)	Alg-User	0.038 ± 0.492	0.064 ± 0.607
	User-User	0.352 ± 0.398	0.275 ± 0.546
Berlin (sentiments)	Alg-User	0.157 ± 0.415	0.073 ± 0.459
	User-User	0.187 ± 0.218	0.235 ± 0.613
Krakow (opinion)	Alg-User	0.285 ± 0.427	0.319 ± 0.507
	User-User	0.411 ± 0.414	0.436 ± 0.553
Warsaw (opinion)	Alg-User	0.440 ± 0.378	0.429 ± 0.399
	User-User	0.160 ± 0.488	0.155 ± 0.573
Berlin (opinion)	Alg-User	0.380 ± 0.358	0.314 ± 0.521
	User-User	0.333 ± 0.563	0.248 ± 0.630
All sentiments	Alg-User	0.213 ± 0.436	0.257 ± 0.506
	User-User	0.226 ± 0.382	0.245 ± 0.505
All opinions	Alg-User	0.347 ± 0.400	0.345 ± 0.477
	User-User	0.324 ± 0.469	0.316 ± 0.573
All	Alg-User	0.287 ± 0.419	0.306 ± 0.489
	User-User	0.285 ± 0.436	0.288 ± 0.544

Table 3 Factors that influence the human evaluator

Question	*Mean*	*Std.*	*Median*
I give lower ratings to comments with many typos	–0.531	1.174	–1
I give lower ratings to comments written in bad English	–0.347	1.251	–1
I give higher ratings to well-thought comments (the comments where people discuss what is so unique in the picture instead of just saying that the photo is good)	1.102	1.141	1
I give higher ratings to comments with many exclamation marks	–0.918	1.037	–1
I give higher ratings to comments if I encounter some type of words (among all possible) that relate to sentiment/opinion expressions	0.796	0.865	1
I weigh equally all adjectives with positive meaning Example: There is no perceptual difference between the two sentences (1) Beautiful place and (2) moving environment	0	1.099	0
I weigh equally all adjectives with negative meaning Example: there is no perceptual difference between the two sentences (1) Ugly place and (2) sad place	–0.224	1.104	0
My rating decision was influenced by the overall number of comments for a particular image	–0.449	1.081	–1

5.3 Results and discussion

Table 2 shows the average results of the algorithm-user and user-user rank and score agreements combined with standard deviation. In the case of Auschwitz and Warsaw, the rank and score agreements between the algorithm and the users are considerably higher than the agreement between users. In all other cases except for Dachau, the level of agreement is similar between the algorithm and users. We can observe a notably big difference between the algorithm and the users for the Dachau region where the user-user rank and score agreements are higher. However, the significance test (denoted as p-value) shows no evidence for statistical difference in all cases. Table 3 shows the answers of participants to eight questions.

The difficulties users experienced, e.g., completing the task and working with different interpretations, are reflected only on a moderate level of user-user agreement on the same comment sets. This tendency shows that for both opinions and sentiments criteria, the users' level of opinion is more similar to the algorithm than the level of agreement among the users. The fact that the user-algorithm agreement is about the same as the user-user agreement is a strong support for the algorithmic approach. It could not be expected that a user-algorithm agreement would exceed the user-user agreement in such a difficult task. The conclusion that can be drawn is that the algorithm in essence is equal to or as good as an average human user, which is promising.

As may be expected, the user-algorithm agreement is generally higher on opinions than on sentiments. As mentioned in Section 4.4.4, the algorithmic separation has a slight tendency to misclassify opinions as sentiments. While some opinions might be missed, the opinion score remains unaffected by falsely regarded sentiments and thus remains accurate.

The opinion analysis for the different regions Krakow, Warsaw and Berlin worked quite well. The sentiment analysis, in contrast, is more heterogeneous. While the algorithm worked well for Auschwitz, the results were less convincing for Berlin and especially for Dachau. A deeper investigation revealed the cause. Apparently,

the comparatively low user-algorithm agreement in the Dachau Region was strongly influenced by one document set, in which three users heavily agreed in disagreeing with the algorithmic result. The user-algorithm agreement was –0.667 (IRA) and –0.65 (ICC), while the user-user agreement was 0.733 (IRA) and 0.985 (ICC). With the purpose of learning about the reasons for this strong deviation in agreement, a further analysis was conducted. Interestingly, all users rated the top-ranked comment file as last and the second as penultimate, which was the main cause for the extreme user-algorithm disagreement. The respective algorithmically top-ranked comment file included many more comments than the algorithmically low-ranked ones. While the latter ones each contained only one sentence expressing negative sentiment and no opinion at all, the two top-ranked comments contained both many positive POs and many very negative sentiments. While our algorithm is tuned to ignore opinions when evaluating sentiments, the users in this case apparently behaved differently, as revealed by some of their answers.

One of the three users answered in the questionnaire that she agreed with the ranking order of the algorithm, despite ranking quite differently herself. The same user also agreed that her rating had been influenced by the overall number of comments in the comment file. It seems she down-ranked the files with more comments. The second user disagreed with the algorithmic ranking, but did not reveal any further details. The third one strongly disagreed with the algorithmic ranking order and stated not to have been influenced by the number of comments. However, her textual explanation was interesting:

> "I looked at the sentiments expressed to determine if they were positive or negative. I did not take into account grammar or punctuation, I looked at what the comments had to say. Even though the one comment file had lots of comments, I felt many of them were more positive or actually opinions of the photo so I said this was the photo with the most positive sentiments contrary of what the algorithm concluded."

Apparently her decision had been influenced by the large number of positive opinions, which somehow attenuated the impression of the negative sentiments.

Thus, it can be concluded that this case reflects weaknesses of the user study rather than the weaknesses of the algorithm.

5.3.1 Limitations

Photo comments may be quite long and each photo may have many comments. Memorising several comment texts with respect to certain criteria and evaluating them in relation to each other is demanding. Therefore, we found five comment text files to be a good trade-off between providing the user with enough data to make his/her reply meaningful and at the same time not to overburden the evaluator. It is also not practical to ask users for real-valued scores without providing them with a sound basis for their decisions, since we did not want them to simulate any kind of algorithmic behaviour. Even the mere ranking of only a limited number of comments does not seem to be a trivial task. According to the users' remarks, differentiating between sentiments and opinion was especially difficult.

We tried to minimise the potential bias introduced by the outlined problems by designing the user task as simple and clear as possible. The drawback of giving only a small set of comments to each user was reduced by averaging over many different sets. Still, in the Dachau region one of the randomly drawn sets considerably influenced

the overall result for the whole region. In addition, the users apparently had varying notions of opinions and sentiments. We tried to prevent this by carefully explaining the differences and providing a large list of examples at the beginning of the study. However, the results reflected that some people did not perceive negative sentiments that strongly if they were coupled with positive opinions.

5.3.2 Additional insights from the questionnaire

In addition to the ranking, we requested the users to fill out a questionnaire in order to learn more about human behaviour when rating opinions or sentiments. The results show that users do not have the tendency to give lower ratings to comments if they contain many typos or are written in bad English. This is consistent with the behaviour of our algorithm, which does not provide special treatment for those cases, unless an opinion/sentiment word or photo feature could not be detected because of a typo. Additionally, users do not give higher ratings to comments with many exclamation marks, which are also ignored in our algorithm. Similar to our algorithm and to our expectation, users take the occurrence of sentiment and opinion words into account and do not tend to weigh them equal. In addition, most users declared that they had not been influenced by the overall number of comments for a particular image, which is the only point where user behaviour deviates from the scoring strategy used by the algorithm. To a certain extent, our algorithm tends to give higher ratings to photos with larger number of comments since the opinion and sentiment scores are calculated based on a sentence, and then added up. In the case of online photo collections, this makes sense as more comments show a higher interest in the photo. Yet, the sentence by sentence score combination can be easily changed to a different strategy. For example by picking only one sentence that has the highest opinion or sentiment strength.

6 Conclusions

This paper introduces a practical unsupervised approach to the analysis of opinions in photo comments. Our approach is capable of identifying two types of opinions from the comments: opinions that are related to the quality of the photo and GSs or moods expressed towards the objects shown on the photo. Unlike most of the existing approaches in which binary (negative or positive) opinion orientation is used, we model opinion orientation using a real-valued scale.

Using linguistic features, we built a finite lexicon of adjectives and calculated their opinion strength using a word importance paradigm borrowed from the information retrieval field combined with the concepts of Zipf's least effort, regularity in word usage, and semantic differentiation of Osgood. The opinion orientation (negative or positive sign) is calculated using a predefined lexicon of positive and negative opinion-bearing words. The identification and separation of POs is based on a semi-automatic method for photo feature extraction and a set of predefined syntactic opinion reference patterns. We applied the cumulative sum to calculate the overall opinion and sentiment scores of comments. This allows a dynamic update of scores if new comments are added to the photo. However, other strategies for overall opinion and sentiment scores can be easily applied.

We conducted a user study in which we analysed factors that influence the human evaluator during the ranking of photo comments. Our study included 49 participants,

who evaluated photo comments from five different regions across Central Europe on a predefined criterion (opinions or sentiments). The results of the user study showed that there is a high variability in agreements between participants themselves, and between the algorithm and the users. However, there was no statistical significance to the difference between the algorithm and the participants, which allows us to conclude that the performance of our algorithm is comparable to the performance of the average user.

The approach is potentially useful in other domains where different kinds of opinions have to be separated. One popular example are movie reviews where one aspect in comments is the plot of the movie and another aspect is the opinion about the movie. For example, a character might be identified as being evil, while the actor does a good job of embodying it. Hence, opinion words relating to the plot should be separated from user opinions.

In our future work, we shall perform a thorough comparison of the pattern-based approach with standard methods and the analysis of photo comments that contain languages other than English. In addition, we shall work on the improvement of the score assignment algorithm taking into account additional factors that were revealed during the user study.

Acknowledgements

This work was partially funded by the German Research Society (DFG) under grant GK-1042 (research training group 'Explorative analysis and visualisation of large information spaces') and by the research initiative 'Computational analysis of linguistic development' (CALD) at the University of Konstanz, Germany.

References

Argamon, S., Bloom, K., Esuli, A. and Sebastiani, F. (2007) 'Automatically determining attitude type and force for sentiment analysis', *Proceedings of the 3rd Language and Technology Conference*, pp.369–373.

Chesley, P., Vincent, B., Xu, L. and Srihari, R. (2006) 'Using verbs and adjectives to automatically classify blog sentiment', *Training*, Vol. 580, No. 263, p.233.

Das, S. and Chen, M. (2007) 'Yahoo! for Amazon: sentiment extraction from small talk on the web', *Management Science*, Vol. 53, No. 9, pp.1375–1388.

Dave, K., Lawrence, S. and Pennock, D. (2003) 'Mining the peanut gallery: opinion extraction and semantic classification of product reviews', in *Proceedings of the 12th International Conference on World Wide Web*, p.528.

Ding, X., Liu, B. and Yu, P. (2008) 'A holistic lexicon-based approach to opinion mining', in *Proceedings of the International Conference on Web Search and Web Data Mining*, pp.231–240.

Drake, A., Ringger, E. and Ventura, D. (2008) 'Sentiment regression: using real-valued scores to summarize overall document sentiment', in *2008 IEEE International Conference on Semantic Computing*, pp.152–157.

Esuli, A. and Sebastiani, F. (2006) 'SentiWordNet: a publicly available lexical resource for opinion mining', in *Proceedings of LREC*, Vol. 6, pp.417–422.

Fahrni, A. and Klenner, M. (2008) 'Old wine or warm beer: target-specific sentiment analysis of adjectives', in *AISB 2008 Convention Communication, Interaction and Social Intelligence*, Vol. 1, p.60.

Fellbaum, C. (1998) *WordNet An Electronic Lexical Database*, MIT press Cambridge, MA.

Gamon, M. (2004) 'Sentiment classification on customer feedback data: noisy data, large feature vectors and the role of linguistic analysis', in *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Hassan, A. and Radev, D. (2010) 'Identifying text polarity using random walks', in *ACL 10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Morristown, NJ, USA, pp.395–403.

Hatzivassiloglou, V. and McKeown, K. (1997) 'Predicting the semantic orientation of adjectives', in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, p.181.

Jijkoun, V., de Rijke, M. and Weerkamp, W. (2010) 'Generating focused topic specific sentiment lexicons', in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp.585–594.

Jones, K.S. (1972) 'A statistical interpretation of term specificity and its application in retrieval', *Journal of Documentation*, Vol. 28, No. 1, pp.11–21.

Kamps, J., Marx, M., Mokken, R. and De Rijke, M. (2004) 'Using WordNet to measure semantic orientation of adjectives', in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Vol. 4, pp.1115–1118.

Kisilevich, S., Rohrdantz, C. and Keim, D. (2010) 'Beautiful picture of an ugly place. Exploring photo collections using opinion and sentiment analysis of user comments', in *Computational Linguistics & Applications (CLA 10)*, pp.419–428.

Liu, B. (2009) 'Sentiment analysis and subjectivity', in N. Indurkhya and F.J. Damerau (Eds.): *Handbook of Natural Language Processing*.

Oelke, D., Hao, M., Rohrdantz, C., Keim, D.A., Dayal, U., Haug, L-E. and Janetzko, H. (2009) 'Visual opinion analysis of customer feedback data', in *VAST 09 Proceedings of the 2009 IEEE Symposium on Visual Analytics Science and Technology*, pp.187–194.

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C. and Smeaton, A. (2009) 'Topic-dependent sentiment analysis of financial blogs', in *Proceeding of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*, pp.9–16.

Osgood, C. (1957) *The Measurement of Meaning*, University of Illinois Press, Champaign, Illinois.

Popescu, A-M. and Etzioni, O. (2005) 'Extracting product features and opinions from reviews', in *HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Morristown, NJ, USA, pp.339–346.

Qiu, G., Liu, B., Bu, J. and Chen, C. (2009) 'Expanding domain sentiment lexicon through double propagation', in *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pp.1199–1204.

Riloff, E. and Wiebe, J. (2003) 'Learning extraction patterns for subjective expressions', in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pp.105–112.

Sahlgren, M., Karlgren, J. and Eriksson, G. (2007) 'SICS: valence annotation based on seeds in word space', in *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp.296–299.

Salton, G. and Buckley, C. (1988) 'Term-weighting approaches in automatic text retrieval', *Information Processing & Management*, Vol. 24, No. 5, pp.513–523.

Salvetti, F., Lewis, S. and Reichenbach, C. (2004) 'Automatic opinion polarity classification of movie reviews', *Colorado Research in Linguistics*, Vol. 17, No. 1.

Shrout, P. and Fleiss, J. (1979) 'Intraclass correlations: uses in assessing rater reliability', *Psychol. Bull.*, Vol. 86, No. 2, pp.420–428.

Subrahmanian, V. and Reforgiato, D. (2008) 'AVA: adjective-verb-adverb combinations for sentiment analysis', *IEEE Intelligent Systems*, Vol. 23, No. 4, pp.43–50.

Toutanova, K. and Manning, C. (2000) 'Enriching the knowledge sources used in a maximum entropy part-of-speech tagger', in *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Association for Computational Linguistics, pp.63–70.

Turney, P. (2002) 'Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp.417–424.

Wiebe, J. (2000) 'Learning subjective adjectives from corpora', in *Proceedings of the National Conference on Artificial Intelligence*, pp.735–741.

Wiebe, J., Brace, R. and O'Hara, T. (1999) 'Development and use of a gold-standard data set for subjectivity classifications', in *Annual Meeting-Association for Computational Linguistics*, Vol. 37, pp.246–253.

Zipf, G. (1949) *Human Behavior and the Principle of Least effort An Introduction to Human Ecology*, Addison-Wesley Press, Boston, Massachusetts.

Notes

1 For the complete list of 20 most frequent adjectives in the five regions please refer to Kisilevich et al. (2010).

2 The Used Part-of-Speech Tags Follow the Penn Treebank Tag-set Definition, available at http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html.