Opinion Mining on the Web 2.0 – Characteristics of User Generated Content and Their Impacts

Gerald Petz¹, Michał Karpowicz¹, Harald Fürschuß¹, Andreas Auinger¹, Václav Střítešský², and Andreas Holzinger³

¹ University of Applied Sciences Upper Austria, Campus Steyr, Austria
(gerald.petz,michal.karpowicz,harald.fuerschuss,
andreas.auinger)@fh-steyr.at
² University of Economics, Prague
stritesv@vse.cz
³ Medical University Graz, Medical Informatics, Statistics and Documentation, Austria
andreas.holzinger@medunigraz.at

Abstract. The field of opinion mining provides a multitude of methods and techniques to be utilized to find, extract and analyze subjective information, such as the one found on social media channels. Because of the differences between these channels as well as their unique characteristics, not all approaches are suitable for each source; there is no “one-size-fits-all” approach. This paper aims at identifying and determining these differences and characteristics by performing an empirical analysis as a basis for a discussion which opinion mining approach seems to be applicable to which social media channel.

Keywords: opinion mining, user generated content, sentiment analysis, text mining, content extraction, language detection, Internet slang, text mining.

1 Introduction and Motivation for Research

Opinion mining (some authors use “sentiment analysis” synonymously), deals with analyzing people’s opinions, sentiments, attitudes and emotions towards different brands, companies, products and even individuals [1], [2]. The rise of the Web 2.0 and its user generated content led to many changes of the Internet and its usage, as well as a change in the communication processes. The user created content on the Web 2.0 can contain a variety of important market research information and opinions, through which economic opportunities as well as risks can be recognized at an early stage. Some of the challenges for qualitative market research on the Web 2.0 are on the one hand the variety of information and on the other hand the huge amount of rapidly growing and changing data.

Besides the typical challenges known from natural language processing and text processing, many challenges for opinion mining in social media sources make the detection and processing of opinions a complicated task:

© Springer-Verlag Berlin Heidelberg 2013
Noisy texts: User generated contents in social media tend to be less grammatically correct, they are informally written and have spelling mistakes. These texts often make use of emoticons and abbreviations or unorthodox capitalisation [3], [4].

Language variations: Texts in user generated content typically contain irony and sarcasm; texts lack contextual information but have implicit knowledge about a specific topic [5].

Relevance and boilerplate: Relevant content on webpages is usually surrounded by irrelevant elements like advertisements, navigational components or previews of other articles; discussions and comment threads can divert to non-relevant topics [5–7].

Target identification: Search-based approaches to opinion mining often face the problem that the topic of the retrieved document does not necessarily match the mentioned object [5].

In the field of opinion mining, where language-specific tools, algorithms and models are frequently utilized, these challenges have quite an important impact on the properness of results, since the application of improper methods leads to incorrect or worse sentiment analysis results.

1.1 Objective and Methodology

The objective of this paper is to investigate the differences between social media channels and to discuss the impacts of their characteristics to opinion mining approaches. To attain this objective, we set up a methodology as follows:

(i) In the first step, we identify the most popular approaches for opinion mining in the scientific field and their underlying principles of detecting and analyzing text.

(ii) As a second step we identify and deduce criteria from literature to exhibit differences between the different kinds of social media sources regarding possible impacts on the quality of opinion mining.

(iii) Subsequently, we carry out an empirical analysis based on the deduced criteria in order to determine the differences between several social media channels. The social media channels taken into consideration in the third step are: social network services (Facebook), microblogs (Twitter), comments on weblogs and product reviews (Amazon and other product review sites).

(iv) In the last step, the social media source types need to be correlated with applicable opinion mining approaches based on their respective characteristics.

The next section gives a short overview about related work and approaches of opinion mining; section 3 describes the empirical analysis and discusses impacts of the characteristics of user generated content to opinion mining.
2 Background, Related Work

Opinion mining deals with different methods and algorithms from computational linguistics and natural language processing in order to find, extract and analyze people’s opinions about certain topics.

2.1 Opinion Definition

Liu defines an opinion as a quintuple \((e_i, a_{ij}, s_{ijkl}, h_k, t_l)\), where \(e_i\) is the name of an entity, \(a_{ij}\) is an aspect of \(e_i\), \(s_{ijkl}\) is the sentiment on aspect \(a_{ij}\) of entity \(e_i\), \(h_k\) is the opinion holder and \(t_l\) is the time, when the opinion is expressed. An entity is the target object of an opinion; it is a product, service, topic, person, or event. The aspects represent parts or attributes of an entity (part-of-relation). The sentiment is positive, negative or neutral or can be expressed with intensity levels. The indices \(i, j, k, l\) indicate that the items in the definition must correspond to one another [1].

2.2 Main Research Directions and Technical Approaches

Several main research directions can be identified [2], [8]: (1) Sentiment classification: The main focus of this research direction is the classification of content according to its sentiment about opinion targets; (2) Feature-based opinion mining (or aspect-based opinion mining) is about analysis of sentiment regarding certain properties of objects (e.g. [9], [10]); (3) Comparison-based opinion mining deals with texts in which comparisons of similar objects are made (e.g. [11]).

Opinion mining has been investigated mainly at three different levels: document level, sentence level and entity/aspect-level. Most classification methods are based on the identification of opinion words or phrases. The underlying algorithms can be categorized as follows: (1) Supervised learning (e.g. [12], [13]), (2) Unsupervised learning (e.g. [14]), (3) Partially supervised learning (e.g. [15]), (4) Other approaches / algorithms like latent variable models (hidden Markov model HMM [16]), conditional random fields CRF [17]), latent semantic association [18], pointwise mutual information (PMI) [19].

Due to the amount of different techniques, several researchers experimented with different algorithms and drew comparisons between them: [20–22].

2.3 Opinion Mining and Web 2.0

A couple of research papers focus explicitly on Web 2.0: A considerably amount of research work covers weblogs, e.g. [23–26], but most of them investigate the correlation between blog posts and “real life”-situations. Only a few papers evaluate techniques for opinion mining in the context of weblogs; there is no main direction of used techniques. Liu et al. [27] compare different linguistic features for blog sentiment classification, [28] experimented with lexical and sentiment features and different learning algorithms for identifying opinionated blogs. Surprisingly, little research work can be found about opinion mining in the area of discussion forums.
However, microblogs – in particular Twitter – seem to be quite attractive to researchers and a variety of papers focusing on microblogs have been published, e.g. [31–35]. The researchers mainly use supervised learning or semi-supervised learning as the dominant approach to mine opinions on microblogs. Despite the popularity of social network services like Facebook, relatively little research work about opinion mining in social networks can be found (e.g. [36], [37]). There are numerous research papers that deal with product reviews, and there is not one specific approach that seems to perform best. Many authors use text classification algorithms like SVM or Naïve Bayes and combine different techniques to increase the quality of opinion mining results. A promising technique could be LDA (e.g. [38], [39]). [40] proposed an LDA-based model that jointly identifies aspects and sentiments. This model (also e.g. the approach of [41], [42]) assumes that all of the words in a sentence cover one single topic.

3 Research Work and Results

We conducted an empirical analysis in order to find differences between social media channels. The following section describes the empirical analysis as well as the impacts of user generated content on opinion mining.

3.1 Empirical Analysis

Methodology of Survey. When starting the empirical analysis, it lends itself to asking the question of how an appropriate sample should be drawn in order to conduct a representative survey. Basically, a random sample is reasonable, but it is actually a challenge to draw a random sample. Therefore, we have decided to draw a sample of self-selected sources and to make a kind of quota sampling. In order to avoid confounders, systematic errors and bias we define the following constraints: we focus on one specific brand / company (in our case: Samsung) and on a specific time period (in our case: between June, 15th 2011 and Jan, 28th 2013) for all sources in social media. Within this time period we conduct a comprehensive survey; if there are too many entries to perform a comprehensive survey, we draw a random sample of the entries. As we do not want to analyze the official postings of the company, we exclude these postings from the analysis. The data sets were labeled manually by four different human labelers. Before the labeling started, we discussed and defined rules for labeling in order to make the labeling consistent among the labelers [11]. The statistical calculations were carried out using SPSS.

The following sources have been surveyed in four different languages: social network service (Facebook; 410 postings), microblog (Twitter; 287 tweets), blog (387 blog posts), discussion forum (417 posts from 4 different forums) and product reviews (433 reviews from Amazon, and two product review pages). The collection of the data was performed manually for the discussion forums and automated using the API (Twitter, Facebook) and a Web-crawler for the other sources (Amazon).
Evaluation Criteria. In order to compare different social media channels, we need to determine indicators. These indicators – shown in table 1 – are derived from two sources: (i) criteria based on simple frequencies from content analysis, and (ii) criteria derived from the definition of opinions (see section 2.1):

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Description</th>
<th>Scale type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>Describes the language used, e.g. English, German, etc.</td>
<td>Nominal</td>
</tr>
<tr>
<td>Number of words</td>
<td>How many words does a posting (e.g. blog posting, Facebook-post, product review, comment, etc.) contain?</td>
<td>Metric</td>
</tr>
<tr>
<td>Number of sentences</td>
<td>How many sentences does a posting contain?</td>
<td>Metric</td>
</tr>
<tr>
<td>Number of Internet slang abbreviations</td>
<td>How many typical Internet slang abbreviations (e.g. LOL, IMO, IMHO …) does the posting contain?</td>
<td>Metric</td>
</tr>
<tr>
<td>Number of emoticons</td>
<td>How many emoticons (e.g. ;-) :-) :-) o …) does the posting contain?</td>
<td>Metric</td>
</tr>
<tr>
<td>Number of incorrect sentences</td>
<td>How many sentences contain grammatical and orthographical mistakes or typos per posting?</td>
<td>Metric</td>
</tr>
<tr>
<td>Subjectivity</td>
<td>Does the posting contain an opinion? Is the posting subjective or objective?</td>
<td>Nominal</td>
</tr>
<tr>
<td>Opinion holder</td>
<td>Is the opinion holder the author of the posting?</td>
<td>Nominal</td>
</tr>
<tr>
<td>Opinion expression</td>
<td>Is the opinion implicitly or explicitly formulated?</td>
<td>Nominal</td>
</tr>
<tr>
<td>Topic-related</td>
<td>Does the posting refer to the headline / overall topic?</td>
<td>Nominal</td>
</tr>
<tr>
<td>Aspect</td>
<td>Does the opinion refer to one or more aspects of the entity?</td>
<td>Nominal</td>
</tr>
</tbody>
</table>

Results of Survey. All in all we analyzed 1934 postings; in the following section we give a short overview on some key findings:

- **Length of postings**: As expected, the length of the postings differs between the social media channels. The average amount of words per posting is highest in product reviews (approx. 119 words), lowest in microblogs (approx. 14 words). Interestingly, the average amount of words per Facebook posting is only 19 words.
- **Emoticons and Internet slang**: Emoticons are widely used across all analyzed social media channels, with approximately every third (Facebook: 27.8%, Twitter: 24.4%, blogs: 27.6%) to fifth (discussion forums: 20.1%, product reviews: 15.5%)
posting containing them. Internet slang is not prominently featured in the analyzed channels, whereby no significant difference between them was detected. While Tweets contain the highest amount of typical abbreviations (20.2% of posting), they only occur in about 12.8% of all discussion forum posts, product reviews and blog comments. Surprisingly, only 8.3% of the analyzed Facebook comments feature Internet slang.

- **Grammatical and orthographical correctness**: Postings across all social media channels contain many grammatical as well as orthographical errors. The error ratio (number of incorrect sentences divided by number of sentences) is highest in Twitter (48.8%), Facebook (42.7%) and discussion forums (42.3%), and lowest in product reviews (37.2%) and blogs (35.4%). The detailed correlations between the variables were tested with Post-Hoc-tests / Bonferroni: product review / Twitter (p=0.002), Twitter / blog (p=0.0).

- **Subjectivity**: Across all analyzed channels 67.8% of the postings were classified as being subjective, as opposed to 18.1% objective ones. The remaining 14.1% of the postings contain both subjective and objective information. While the highest subjectivity can be detected on Twitter (82.9% of all analyzed Tweets), discussion forums not only feature the fewest subjective posts (50.2%) but also the majority of objective ones (35.5%). Many of the postings in discussion forums do not contain an opinion, but questions, solution suggestions and hints how to solve a specific issue. An interesting discovery is the lack of exclusively objective product reviews – nearly two thirds (71.7%) of the analyzed reviews are solely subjective, while one quarter (25.4%) is based on both subjective and objective information. 2.9% of the reviews are rated as being objective. The detailed correlations between the variables were tested with Post-Hoc-tests / Bonferroni: Facebook / discussion forum (p=0.001), Twitter / product review (p=0.0), Twitter / blog (p=0.033), Twitter / discussion forum (p=0.0).

<table>
<thead>
<tr>
<th>Social media channel</th>
<th>Subjective</th>
<th>Objective</th>
<th>Subjective &amp; objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microblog (Twitter)</td>
<td>82.9%</td>
<td>12.8%</td>
<td>4.3%</td>
</tr>
<tr>
<td>Product Review</td>
<td>71.7%</td>
<td>2.9%</td>
<td>25.4%</td>
</tr>
<tr>
<td>Blog</td>
<td>69.3%</td>
<td>19.6%</td>
<td>11.1%</td>
</tr>
<tr>
<td>Social Network (Facebook)</td>
<td>67.3%</td>
<td>26.1%</td>
<td>6.6%</td>
</tr>
<tr>
<td>Discussion forum</td>
<td>50.2%</td>
<td>35.5%</td>
<td>14.3%</td>
</tr>
</tbody>
</table>

- **Aspects and details**: As expected, the social media channels that tend to feature longer postings contain more details on certain aspects of entities. The detailed figures are exhibited in Table 4. While product review postings go into detail (39.6%) and contain aspects as well as opinions on entity-level (27.0%), Twitter and Facebook-postings mainly contain postings on entity-level (56.6%, 65.4%).
Table 3. Opinions about entities and aspects

<table>
<thead>
<tr>
<th>Social media channel</th>
<th>Contains one or more aspects</th>
<th>Does not contain aspects</th>
<th>Contains opinion about entity and aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discussion forum</td>
<td>60.6%</td>
<td>33.1%</td>
<td>6.3%</td>
</tr>
<tr>
<td>Blog</td>
<td>55.3%</td>
<td>39.1%</td>
<td>5.6%</td>
</tr>
<tr>
<td>Microblog (Twitter)</td>
<td>43.4%</td>
<td>56.6%</td>
<td>0%</td>
</tr>
<tr>
<td>Product Review</td>
<td>39.6%</td>
<td>33.4%</td>
<td>27.0%</td>
</tr>
<tr>
<td>Social Network</td>
<td>33.0%</td>
<td>65.4%</td>
<td>1.6%</td>
</tr>
</tbody>
</table>

- **Opinion holder**: The survey exhibited that in most cases the opinion holder is equal to the author of the posting; in Facebook, Twitter, product reviews and blogs between 95% and 97.6% of the postings reveal the author as the opinion holder. Only the postings in the discussion forums have a lower percentage (90.7%). 6.2% of the entries in discussion forums have several opinion holders, and 3.1% depict the opinion of another person.

- **Topic relatedness**: At the beginning of our survey we were curious about the users’ “discipline” regarding the topic relatedness of their postings. Surprisingly, the postings in all the social media channels are highly related to the overall discussion topic. As shown in the following table, the highest relatedness can be found in discussion forums, which may be related to the presence of moderators and forum rules.

Table 4. Topic relatedness

<table>
<thead>
<tr>
<th>Social media channel</th>
<th>Topic related</th>
<th>Not topic related</th>
<th>Topic and non-topic related content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discussion forum</td>
<td>95.6%</td>
<td>3.4%</td>
<td>1.0%</td>
</tr>
<tr>
<td>Microblog (Twitter)</td>
<td>95.3%</td>
<td>4.7%</td>
<td>0%</td>
</tr>
<tr>
<td>Product review</td>
<td>93.1%</td>
<td>1.2%</td>
<td>5.8%</td>
</tr>
<tr>
<td>Blog</td>
<td>92.6%</td>
<td>6.3%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Social Network</td>
<td>82.3%</td>
<td>16.6%</td>
<td>1.1%</td>
</tr>
</tbody>
</table>

**Discussion of Survey.** The criteria we used for the survey are often criticized in research papers for their ambiguity, e.g. subjective vs. objective. The team that conducted the survey exchanged their experiences and carried out multiple evaluations on the same sample set. There remains the question of how to conduct a survey that is both representative and accomplishable with manageable efforts. In our survey we used one brand from the electronic consumer market, but the results may vary depending on other market segments or genres.
3.2 Impact on Opinion Mining

Based on the empirical analysis the following impacts can be derived for the opinion mining process:

**Impacts on Opinion Mining Process.** Many research papers in the field of opinion mining assume grammatically correct texts [4], but as shown in the empirical analysis, user generated texts contain many mistakes, emoticons and Internet slang words. Therefore it is reasonable and necessary to preprocess texts from Web 2.0-sources. In some cases the text languages changed on the same channel, e.g. some Facebook postings on the German Facebook site are written in English, Turkish and other languages. In these cases the application of language detection methods is reasonable. In general, because of the grammatical mistakes, grammar-based approaches (e.g. [44], [45]) are not appropriate.

The above figures showed, that user generated texts contain Internet slang as well as emoticons. These text parts could be considered as input for feature generation to improve sentiment classification. Furthermore, people often use different names for the same object, e.g. “Samsung Galaxy S3” is also being called “Galaxy S3” or “SGS3”, which makes the extraction of entities or aspects more difficult.

**Characteristics and Impacts of Social Media Channels.** The following table gives a short overview about the impacts of each investigated social media channel:

<table>
<thead>
<tr>
<th>Social media channel</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Discussion forum</strong></td>
<td>The empirical analysis revealed, that discussions in forums are often organized in discussion threads, users respond to other user’s questions and comments, and forum postings often contain coreferences – all these factors make opinion mining more difficult and a variety of approaches have to be adopted to discussion forums. More research work is required to evaluate, which methods perform best.</td>
</tr>
<tr>
<td><strong>Microblog (Twitter)</strong></td>
<td>The characteristics of Twitter can be summarized as follows: many grammatical errors, short sentences, heavy usage of hashtags and other abbreviations. That already led researchers to taking Twitter characteristics into consideration, e.g. Davidov et al. [46] use Twitter characteristics and language conventions as features, Zhang et al. [47] combine lexicon-based and learning-based methods for Twitter sentiment analysis. The usage of part-of-speech features does not seem to be useful in the microblogging domain (e.g. [48]).</td>
</tr>
</tbody>
</table>
Several researchers proposed models to identify aspects and sentiments; a few of them assume that all of the words in a sentence cover one single topic. This assumption may be reasonable for product reviews, but this assumption has to be questioned for Facebook, because there are often missing punctuations and it is - even for humans - not easy to detect the boundaries of sentences and to find out the meaning of expressions.

Many research papers that focus on blogs do not unfold how comments to the blog posts are taken into consideration. The comments to blog posts vary in terms of length, coreferences, etc., and thus can be very short answers when the user replies with a short answer or quite long texts when users discuss a topic controversially for instance. From our point of view, depending on the type of the blog (corporate blog vs. j-blog) both the blog posting and the blog comments can be interesting sources for opinion mining.

Because users can interact with each other, respond to questions and the amount of grammatical mistakes, there are similar challenges like with discussion forums. More research work is required.

4 Conclusion and Further Research

This paper discusses the differences of social media channels including microblogs (Twitter), social network services (Facebook), weblogs, discussion forums and product review sites. A survey has been conducted to exhibit the differences of these social media channels, and implications for opinion mining have been derived. The survey covers only the contents related to one specific brand, because the authors wanted to emphasize the viewpoint of a company; of course, the results could be different in other genres (e.g. political discussions), which would require more empirical analysis. The work shows that the dominant approach to mine opinions on microblogs is supervised or semisupervised learning; while for product reviews a wide range of techniques is applied.

Further research work should be conducted: (i) Measure and compare the factual implications of the characteristics of social media on the performance of the different opinion mining approaches, and (ii) conduct more research work on alternative (statistical / mathematical) approaches.

Acknowledgements. This work emerged from the research projects OPMIN 2.0 and SENOMWEB. The project SENOMWEB is funded by the European Regional Development fund (EFRE, Regio 13). OPMIN 2.0 is funded under the program COIN – Cooperation & Innovation. COIN is a joint initiative launched by the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT) and the Austrian Federal Ministry of Economy, Family and Youth (BMWFJ).
References