Information Processing and Management 50 (2014) 899–908

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Computational approaches for mining user's opinions on the Web 2.0

Gerald Petz [a,*], Michał Karpowicz [a], Harald Fürschuß [a], Andreas Auinger [a], Václav Stříteský [b], Andreas Holzinger [c]

[a] University of Applied Sciences Upper Austria, Campus Steyr, Wehrgrabengasse 1-3, 4400 Steyr, Austria
[b] University of Economics, Prague, W. Churchill Sq. 4, 13067 Prague, Czech Republic
[c] Medical University, Graz, Medical Informatics, Statistics and Documentation, Auenbruggerplatz 2/V, 8036 Graz, Austria

ARTICLE INFO

Article history:
Received 14 August 2013
Received in revised form 14 February 2014
Accepted 24 July 2014
Available online 24 August 2014

Keywords:
Opinion mining
Noisy text
Text preprocessing
User generated content
Data mining

ABSTRACT

The emerging research area of opinion mining deals with computational methods in order to find, extract and systematically analyze people's opinions, attitudes and emotions towards certain topics. While providing interesting market research information, the user generated content existing on the Web 2.0 presents numerous challenges regarding systematic analysis, the differences and unique characteristics of the various social media channels being one of them. This article reports on the determination of such particularities, and deduces their impact on text preprocessing and opinion mining algorithms. The effectiveness of different algorithms is evaluated in order to determine their applicability to the various social media channels. Our research shows that text preprocessing algorithms are mandatory for mining opinions on the Web 2.0 and that part of these algorithms are sensitive to errors and mistakes contained in the user generated content.

1. Introduction

Opinion mining deals with analyzing people's opinions, attitudes and emotions towards different brands, companies, products and even individuals (Balahur, 2013; Liu, 2012; Pang & Lee, 2008). Although related research areas to opinion mining such as natural language processing (NLP), information extraction and information retrieval have quite a considerable history, the research on mining people's opinions has become quite popular in the last couple of years with the rise of the Web 2.0. User generated content on the Social Web can contain a variety of relevant market research information and deeply analyzing and exploiting it leads to more targeted business decisions (Guozheng, Faming, Fang, & Jian, 2008; Liu, 2008).

Analyzing opinions on the Social Web is met with a variety of challenges: (i) the "usual" challenges known from natural language processing (such as word sense disambiguation, topic recognition and co-reference resolutions) and (ii) challenges arising from user generated content:

* Corresponding author. Tel.: +43 (0)50804 33410.
E-mail addresses: gerald.petz@fh-steyr.at (G. Petz), michal.karpowicz@fh-steyr.at (M. Karpowicz), harald.fuerschuss@fh-steyr.at (H. Fürschuß), andreas.auinger@fh-steyr.at (A. Auinger), stritesv@vse.cz (V. Stříteský), andreas.holzinger@medunigraz.at (A. Holzinger).

http://dx.doi.org/10.1016/j.ipm.2014.07.005
0306-4573/

- *Noisy texts, language variations*: User generated texts tend to be less grammatically correct and often use specific characters to express emotions (emoticons), abbreviations and unorthodox capitalization. (Abbasi, Chen, & Salem, 2008; Dey & Haque, 2009). Moreover, social media texts typically assume a higher level of knowledge about the context by the reader than more formal texts (Maynard, Bontcheva, & Rout, 2012).
- *Relevance and boilerplate*: When web texts and social media texts are gathered using a web crawler, the gained texts usually contain irrelevant content like advertisements, navigational elements or previews of other articles (Maynard et al., 2012; Petz et al., 2012; Yi & Liu, 2003).
- *Target identification*: Search-based approaches have to deal with the problem, that topics of retrieved documents do not necessarily match the mentioned sentiment object (Maynard et al., 2012).
- *Big data challenges*: That can be broken into several contexts such as temporal, spatial and spatio-temporal contexts (Derczynski, Yang, et al., 2013; Maynard, Dupplaw, & Hare, 2013).

Due to these challenges, research papers usually deal with assumptions and constraints: Many of the approaches to analyze opinions assume linguistically correct texts (Dey & Haque, 2009), others focus on specific social media resources (e.g. *Twitter* as a basis for opinion mining Bollen, Mao, & Zeng, 2011; Davidov, Tsur, & Rappoport, 2010; Pak & Paroubek, 2010; or newswire text Balahur, Steinberger, van der Goot, Pouliquen, & Kabadjov, 2009; Sayeed, 2011; or Blogs Leshed & Kaye, 2006; Mishne & Glance, 2006; Zhang, Yu, & Meng, 2007). The utilization of text preprocessing steps prior to sentiment analysis approaches is quite important in order to achieve good results.

The objectives of this paper are (i) to investigate the differences between social media channels regarding opinion mining and (ii) to evaluate the effectiveness of various text preprocessing algorithms as a subtask of opinion mining in these social media channels. To attain these objectives, we set up the research methodology as follows:

(1) Identification of popular approaches and algorithms to carry out text preprocessing as a prior step to sentiment analysis.
(2) Identification of differences between social media channels and deduction of impacts on opinion mining and text preprocessing.
(3) Evaluation of the effectiveness and properness of several algorithms in order to determine their applicability.

The rest of the paper is organized as follows: in the next section we discuss some related work in the field of opinion mining. We then report in Section 3 on the characteristics of user generated content in different social media channels. Section 4 discusses the impacts of these characteristics on some frequently used algorithms and evaluates their performance regarding noisy text.

2. Related work, background

2.1. Sentiment analysis and opinion mining

Pang and Lee (2008) and Liu (2012) present a detailed review of opinion mining. 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 numeric scores (such as star-ratings). The indices i, j, k, l indicate that the items in the definition must correspond to one another. (Liu, 2012; Wilson, Wiebe, & Hoffmann, 2009)

There are several main research directions (Kaiser, 2009; Pang & Lee, 2008): (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* Hu & Liu, 2004b; Liu, Hu, & Cheng, 2005) is about analysis of sentiment regarding certain properties of objects (e.g. Hu & Liu, 2004a); (3) *comparison-based opinion mining* deals with texts in which comparisons of similar objects are made (e.g. Jindal & Liu, 2006a, 2006b). Other research directions focus on multilingual opinion mining (e.g. Banea, Mihalcea, & Wiebe, 2010; Steinberger, Lenkova, Kabadjov, Steinberger, & Goot van der, 2011) and on cross-domain sentiment analysis (e.g. Bollegala, Weir, & Carroll, 2011; Pan, Ni, Sun, Yang, & Chen, 2010).

The classification of texts regarding sentiment polarity can be done at three different levels: (1) document level, (2) sentence level and (3) entity and aspect-level. There are several approaches to analyze opinions: (1) corpus-based approaches (e.g. Hatzivassiloglou & Wiebe, 2000; Turney, 2002; Wiebe & Mihalcea, 2006) and dictionary-based/lexicon-based approaches (e.g. Ding, Liu, & Yu, 2008; Hu & Liu, 2004a; Kim & Hovy, 2004; Popescu & Etzioni, 2005; Steinberger et al., 2012), (2) machine learning approaches. These approaches can be categorized as follows:

(1) *Supervised learning*: Supervised learning ("classification") is a machine learning task of inferring a function from labeled training data, where statistical methods are applied to construct prediction rules. This type of learning is widely used in real-world applications. Typical supervised learning algorithms are Naïve bayes classifiers, maximum entropy, support vector machines (SVM) and K-Nearest neighbor learning, amongst others (Liu, 2008; Zhang, 2010).

(2) *Unsupervised learning*: Unsupervised learning is often used when the user wants to find hidden structures in unlabeled data and is often called "clustering". A variety of algorithms exist in this subject: k-means, mixture models, hierarchical clustering, etc. (Liu, 2008).
(3) *Other approaches/algorithms*: Since supervised learning needs a large number of labeled data for training and this task is often done manually and therefore is particularly time consuming, some researchers developed partially supervised learning approaches. The tasks include learning from labeled and unlabeled examples ("LU learning") and learning from positive and unlabeled examples ("PU learning"). Exemplary algorithms include Expectation–Maximization (EM) algorithms, co-training and transductive support vector machines (Liu, 2008). One can identify several other algorithms used in opinion mining. E.g. aspect based opinion mining focuses on feature based approaches (e.g. Hu & Liu, 2004a; Liu et al., 2005; Moghaddam & Ester, 2010), while others apply latent variable models like the hidden Markov model (HMM, e.g. Jin, Ho, & Srihari, 2009; Wong, Bing, & Lam, 2011), conditional random fields (CRF, e.g. Choi & Cardie, 2010; Li et al., 2010), latent semantic association (e.g. Guo, Zhu, Guo, & Su, 2011; Guo, Zhu, Guo, Zahng, & Su, 2009; Hofmann, 2001) up to combinations and variations of existing algorithms (Airoldi, Bai, & Padman, 2006; Choi, Cardie, Riloff, & Patwardhan, 2005; Jakob & Gurevych, 2010; Nakagawa, Inui, & Kurohashi, 2010). One important part of the preprocessing steps in opinion mining is Part-of-Speech (POS) tagging; a variety of algorithms can be applied to implement this task: rule based approaches, Markov model approaches, maximum entropy approaches, etc. The majority of research work has focused on POS tagging in English, although there are research efforts under way to extend POS tagging to other languages as well (Güngör, 2010).

Due to the number of different techniques, several researchers experimented with different algorithms and drew comparisons between them: (Chaovalit & Zhou, 2005; Cui, Mittal, & Datar, 2006; Moghaddam & Ester, 2012). Serveral researchers experiment with adaption of machine learning approaches to other languages and domains, e.g. (Boyd-Graber & Resnik, 2010), (Balahur & Turchi, 2013).

2.2. Preprocessing noisy text

Only little research work can be identified on preprocessing of noisy texts. The objective of preprocessing is to produce clean texts for further analysis processes. The tasks usually include identifying and correcting spelling errors, eliminating arbitrary sequences of whitespaces between words, detecting sentence boundaries, eliminating arbitrary use of punctuation marks and capitalization and are usually executed in a pipeline. One of the first research efforts in this area is from Kerninghan et al.; the authors describe a software program that corrects spelling mistakes and typos based on a noisy channel model using Bayesian algorithms (Kemighan, Church, & Gale, 1990). Mikheev presents an approach to detect and correct sentence boundary disambiguation, word disambiguation in terms of capitalization and identification of abbreviations (Mikheev, 2002). A couple of authors used machine learning approaches to deal with noisy texts (e.g. Clark, 2003; Gotoh & Renals, 2000; Kiss & Strunk, 2006). The cleaning approaches are seldom evaluated regarding Web texts or user generated texts in social media. Dey/Haque propose a framework for opinion extraction and mining from noisy text; the framework includes sentence boundary detection, improper case correction and context-dependent spelling correction. In order to determine opinions, the framework uses the cleaned text as input to POS tagging, dependency trees and other algorithms (Dey & Haque, 2009). Derczynski et al. focus on POS tagging for tweets and evaluate the performance of several widely used taggers (e.g. Brant's TnT-tagger, Brill's tagger TBL, Stanford tagger, etc.) that were trained with WSJ-corpus. Not surprisingly, the authors report a weak performance of these taggers when applied to tweets. The most frequent POS tagging errors result from internet slang words, common misspellings, genre-oriented phrases and unknown words. Based on this error analysis, the authors developed a POS tagger, which achieves significantly better results than the traditional POS taggers (Derczynski, Ritter, et al., 2013). In (Derczynski, Maynard, et al., 2013) the authors investigate the impact of the specific characteristics of tweets on several text preprocessing steps: language identification, tokenization, POS tagging and named entity recognition. All in all, the performance of machine learning methods suffers from noisy texts. Normalization approaches (basic resp. strong normalization) of noisy texts offer only small improvements, but are helpful and have to be further developed.

3. Comparison of social media channels

Kessler/Nicolov collected 194 blog entries about cars and digital cameras and calculated some opinion mining relevant statistics (e.g. number of tokens between a sentiment expression and its target; target in the same sentence as their sentiment expression, etc.). (Kessler & Nicolov, 2009) We want to extend these statistics and find differences between social media channels in order to derive impacts on the opinion mining process.

3.1. Methodology

We carried out an empirical analysis of social media texts; for this purpose we collected data from social network services (Facebook; 410 postings), microblogs (Twitter; 287 tweets), blogs (387 blog posts), discussion forums (417 posts from 4 different forums) and product review portals (433 reviews from Amazon, and two product review pages) in four different

languages (English, German, Czech, Polish). In order to conduct a representative survey, we drew a quota sample and therefore focused on one specific brand (Samsung) in a specific time period (between 15 June 2011 and 28 January 2013) in the different social media channels. The collection was performed both manually (discussion forums) and automatically using APIs and Web crawlers. The data sets were labeled manually by four different human labelers. Rules for labeling have been discussed and defined to make the labeling as consistent as possible. We defined several measure criteria (e.g. number of words per posting, number of sentences per posting, etc.) along with their scale type. All in all 1934 postings have been analyzed; the statistical calculations were carried out using SPSS, correlations between variables were tested with Post-Hoc-tests/Bonferroni (Petz et al., 2013).

3.2. Results

In the following section we give a short overview on some key findings

- *Basic figures*: The survey revealed some findings that were to be expected. E.g. the length of the postings differs between the social media channels: the average number of words in a posting is highest in product reviews, and lowest in microblogs (see Table 1).
- *Subjectivity*: Many papers assume that content in social media is highly subjective and therefore a perfect foundation for sentiment analysis tasks. Our analysis revealed that microblogs contain the most subjective information (82.9% of the postings). Product reviews consist of subjective and objective texts; e.g. consumers write both about feelings and impressions about a product and usually summarize facts as well. Surprisingly only about 50% of discussion forums postings contain subjective texts. A closer glance at the discussions forums showed that users often write about hints, tips and instructions of how to deal with a specific problem. 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 ($p = 0.0$) (see Table 2).
- *Grammatical correctness*: Some authors have taken grammatically false texts into consideration, but the vast majority of research work assumes grammatically correct texts. As the following chart exhibits, all social media channels contain many grammatical and orthographical errors. We calculated an error ratio as follows: the number of incorrect sentences divided by the number of sentences. Fig. 1 exhibits the error ratio in more detail; the correlations between the variables were tested with Post-Hoc-tests/Bonferroni: product review/Twitter ($p = 0.002$), Twitter/blog ($p = 0.0$).
- The evaluation of grammatical correctness was not easy: people sometimes do not use punctuation marks, which leads to discussions among the labelers of how many sentences are expressed in that posting; in German texts one can find complete texts without capitalization. Some of the tweets in Twitter are nearly impossible to understand without context knowledge (e.g. "Swagger Shout Out ! @TWAMBIT @MixtapeOficial @SamsungAT @greentravel @__jitesh__ #Sunglasses #ShoutOut"). Additionally, a considerable number of authors lengthen words in order to emphasize specific aspects or feelings (e.g. "Good phone gooooooooooooood").
- *Emoticons, abbreviations*: Surprisingly user generated content does not contain many emoticons (such as ";–)" "<3" "–.–") and abbreviations (e.g. "LOL", "IMHO"). The following tables provide an overview about the usage of emoticons and abbreviations: Table 3 shows that about 23% of postings in social networks contain one emoticon; interestingly quite a low number of postings contain more than two emoticons. The situation regarding abbreviations is quite similar. Because of the limited amount of characters per tweet, Twitter contains the most abbreviations (see Table 4). Some research papers discuss the possibility of using "sentiment hashtags" (like #sarcasm, #fear, #anger) as indicators of emotions or sarcastic texts in Twitter; however, our sample did not contain these patterns.
- *Opinion holder*: In most cases the opinion holder is the author of the posting; with the exception of discussion forum entries, between 95% and 97.6% of the postings reveal the author as the opinion holder. In 90.7% of the postings in discussions forums the author is the opinion holder, 6.2% of the entries in discussion forums have several opinion holders, and 3.1% depict the opinion of another person.
- *Differences between languages*: Clearly, there are differences between the languages regarding the length of the postings (when they are not limited). Although there are differences in terms of usage of emoticons and abbreviations as well, there are no figures that have led to completely new findings other than the ones presented above.

Table 1
Length of postings.

	Number of words		Number of sentences		
	Average	Median	Average	Median	N
Product review	118.9	38	8.1	4	433
Discussion forum	53.6	35	3.9	3	417
Blog	30.6	22	2.7	2	387
Social network (Facebook)	18.9	9	2.0	1	410
Microblog (Twitter)	14.0	13	1.6	1	287

Table 2
Subjective texts in postings.

Social media channel	Subjective (%)	Objective (%)	Subjective and objective (%)
Microblog (Twitter)	82.9	12.8	4.3
Product review	71.7	2.9	25.4
Blog	69.3	19.6	11.1
Social network (Facebook)	67.3	26.1	6.6
Discussion forum	50.2	35.5	14.3

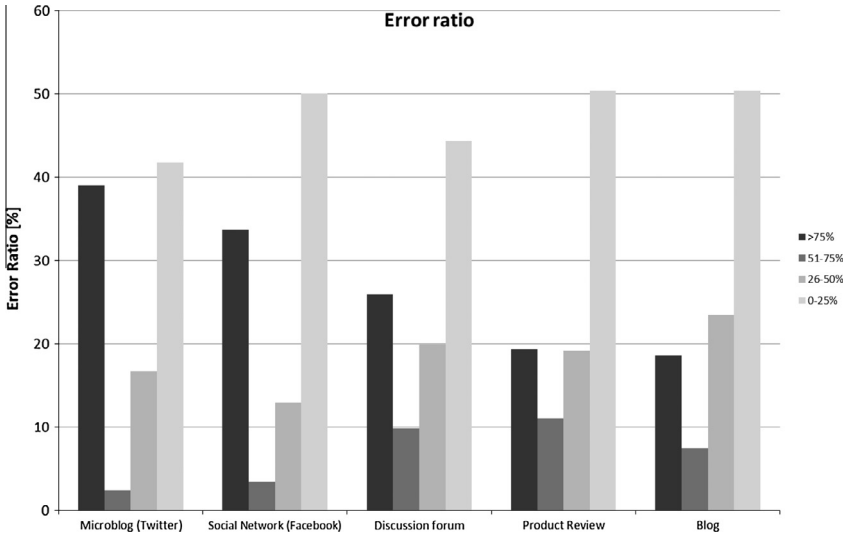

Fig. 1. Error ratio in social media channels.

Table 3
Number of emoticons per posting.

Number of emoticons	1 (%)	2 (%)	>2 (%)
Social network (Facebook)	23.4	3.9	0.5
Blog	21.7	5.2	0.8
Microblog (Twitter)	21.3	2.4	0.7
Discussion forum	15.1	3.6	1.4
Product review	10.4	3.2	1.8

Table 4
Number of abbreviations per posting.

Number of abbreviations	1 (%)	2 (%)	>2 (%)
Microblog (Twitter)	14.6	4.2	1.4
Product review	8.8	1.8	1.8
Blog	8.5	3.6	2.1
Discussion forum	7.4	1.9	2.4
Social network (Facebook)	6.1	1.0	1.2

Since the sample has focused only on one brand in the consumer goods sector, the question arises concerning the extent to which generalizations can be made based on these results and then applied to other areas. We assume that these results can be generalized to other products and brands, but possibly not to political discussions or discussions about arts.

4. Evaluation of Impacts

The text preprocessing is an important step in the opinion mining process. As mentioned above, only few papers deal with noisy texts derived from different social media channels. The following section focuses on the factual impact of the findings above.

4.1. Methodology

We evaluated the factual impact of user generated content on selected algorithms that are frequently used for preprocessing texts during the opinion mining process. The following algorithms were evaluated:

- *Sentence splitting*: Since many other preprocessing steps require text fragments based on sentences, one of the first steps is usually sentence boundary detection. Having little or no language-specific implementations for the Czech and Polish language, only texts in German (422 postings) and English (411 postings) were used. The following algorithms were compared: (i) "ANNIE sentence splitter" and (ii) "Regex splitter" from the software package "GATE" (Version 7.1 build 4485 GATE Cunningham, Maynard, Bontcheva, & Tablan, 2002) and the (iii) "Regex splitter" and from the library "OpenNLP" (version 1.5, model "en-sent.bin" The Apache Software Foundation).
- *Stemming*: Stemming is the process of reducing words to their base form; we used the Snowball stemmer (Porter, 2001) implemented in the Lucene.NET library. Lucene is an open source information retrieval library; Lucene.NET is a port of the library written in C# (The Apache Software Foundation; McCandless, Hatcher, & Gospodnetić, 2010). The German texts were stemmed with the algorithm of Caumanns (Caumanns, 1999) implemented in RapidMiner (Rapid-i).
- *Part of speech tagger*: POS taggers label tokens with their corresponding word type. The implementation of OpenNLP with the English POS model and GATE ANNIE POS tagger (version 7.1 build 4485) were used. These taggers use the Penn Treebank tag set. (The Apache Software Foundation) In order to evaluate performance of POS tagger we used a sample ($n = 150$) of English texts from the above sample set. The texts have been corrected manually by the researchers. Subsequently two taggers have been applied to the texts.
- *Parser*: A parser reveals the structure of sentences, e.g. which words can be grouped into phrases, which words are the subject or the object of a verb. We used the parser from OpenNLP with the model "en-parser-chunking.bin" for the English texts (The Apache Software Foundation).

The evaluation was carried out as follows: a quota sample of the previously collected data was drawn in order to cover each social media channel. Depending on the availability of algorithms for specific languages, we took postings in other languages than English into consideration. Since there is no "gold standard" with labeled training data the researchers corrected the postings (e.g. added punctuation marks, corrected typos, etc. but not editing in order to improve readability or to create meaningful texts). Then the algorithms were compared against these manually corrected postings.

4.2. Results

In the following section we give a short overview on some key findings:

- *Sentence splitting*: Tables 5 and 6 show the percentage of the correct sentence splitting. The performances of the sentence splitters are quite different in the several social media channels; it is relatively accurate for English texts in social networks and product reviews, but comparatively low in blogs and discussions forums. The performance for German texts is worse. A detailed glance at the data reveals, that the sample contains simple sentences for Twitter and Facebook (e.g. "cool", "I so want this camera"), while product reviews, blogs and discussion forums contain a larger number of sentences and higher complexity of the writing. Wrong sentence boundary detection often also arises in that bulleted lists are not interpreted as sentences, while human labelers tend to interpret a bullet item as one (or more) sentences.
- *Stemming*: Due to the simple rules implemented in stemming algorithms, the results of the stemmed texts are comparable. We suspected that typos and strange abbreviations could lead to different stemmed words, but that was not the case in our sample.
- *Part of speech tagging*: The part of speech tagging algorithms seem to be relatively robust regarding noisy texts; the majority of the misjudgments arise because of spelling mistakes. E.g. the sentence "I like ur firm" is tagged as "(I PRP) (like VBP) (ur NN) (firm NN)". If we correct the posting to "I like your firm" the POS tagger annotates as follows: "(I PRP) (like VBP) (your PRP$) (firm NN)". False sentence boundary detection has surprisingly little impact. The high POS output equivalent on the microblogs can be explained in such a way that it is difficult to correct these posts, as there are often many hashtags ("#") and involve direct addressing ("@") and contain only a few words. The following table depicts, whether the output of the POS tagging of the original posting is equivalent to the corrected posting (see Table 7).

Table 5
Percentage of correct sentence splitting (English).

	ANNIE sentence splitter (%)	GATE regex splitter (%)	OpenNLP sentence splitter (%)
Social network (Facebook)	92	92	88
Product review	82	83	74
Microblog (Twitter)	45	45	44
Blog	42	42	43
Discussion forum	77	79	80
Total	82	83	80

Table 6
Percentage of correct sentence splitting (German).

	ANNIE sentence splitter (%)	GATE regex splitter (%)	OpenNLP sentence splitter (%)
Social network (Facebook)	75	10	69
Product review	32	6	31
Microblog (Twitter)	79	7	86
Blog	59	7	62
Discussion forum	64	4	66
Total	59	7	58

- *Parsing*: As expected, parsing is more sensitive to grammar mistakes than is POS tagging. The following table shows, whether the output of the parser of the original posting is grammatically equivalent to the corrected posting. Because of the short sentences in Twitter it is not surprising that the parser output matches best (see Table 8).

4.3. Improvements to the opinion mining process

Some authors suggest stripping symbols in order to retrieve clean texts; as the above figures show, user generated texts contain a lot of emoticons that can express feelings too. If these symbols were removed, additional context information that could be used to improve sentiment analysis would be lost. Dey/Haque propose in their framework several steps for cleaning texts and then use PoS tagging and dependency trees among others for determining the sentiment orientation. The proposed steps – sentence boundary detection, improper case correction and context-dependent spelling corrections – are important and very useful steps, but might not be sufficient for all kinds of social media channels. It is essential for approaches that rely heavily on parsing to execute preprocessing steps, because the output of parsers may vary with the noise included in the input texts.

The characteristics of the different social media channels should be taken more into consideration:

- *Twitter*: Although postings on Twitter have a limited length, the hashtags can deliver some kind of contextual information. Several researchers have already exploited Twitter characteristics, e.g. Davidov et al. (2010) treat hashtags and language conventions as features, Zhang, Ghosh, Dekhil, Hsu, and Liu (2011) combine lexicon-based and learning-based methods for Twitter sentiment analysis. (Brody & Diakopoulos, 2011) suggest detecting lengthened words (e.g. "cooool") because these words are often used as subjective words to emphasize the sentiment they convey. Balahur (2013) suggests a variety of preprocessing steps, including slang replacement, emoticon replacement with their polarity, word normalization, user and topic labeling, and affect word matching. Some researchers query the use of part of speech features in the microblogging domain (e.g. Kouloumpis, Wilson, & Moore, 2011). However, the POS tagging of a grammatically corrected tweet is comparable to the POS tagging of the original tweet.
- *Social networks, blogs, discussion forums, product reviews*: As our survey shows, similar expressions (like "cooool"), abbreviations and emoticons are contained not only in tweets, but in other social media channels as well (with different frequencies). Hence, it seems to be reasonable to adapt the above mentioned techniques to the other channels. Our survey focused on social media channels in four different languages; however, our sample contained several other languages as well. Hence, language detection seems to be reasonable to improve sentiment analysis (Bergsma, McNamee, Bagdouri, Fink, & Wilson, 2012; Petz et al., 2012).

5. Conclusion and further research

Computational approaches on opinion mining are an emerging research topic and there are many challenging future research avenues. In this article we investigated the differences of several social media channels concerning text quality and its factual impacts to a couple of frequently used algorithms in the opinion mining process. In the first section of this work we reported on a survey, which was carried out in order to find the characteristics of social network services, microblogs, blogs, product reviews and discussions forums. The survey focused on the viewpoint of a company; hence the survey covered a specific brand of a manufacturer of electronic devices. In the second section we presented the evaluation of the

Table 7
POS output equivalent.

	OpenNLP (%)	GATE (%)
Microblog (Twitter)	96.7	89.7
Blog	76.7	70.0
Product review	56.7	63.3
Social network (Facebook)	56.7	60.0
Discussion forum	50.0	60.0

Table 8
Comparison of parser output.

	Parser output matches (%)
Microblog (Twitter)	73.3
Blog	56.7
Social network (Facebook)	36.7
Product review	33.3
Discussion forum	13.3

performance in terms of correctness of sentence splitting, stemming, POS tagging and parsing. The results demonstrated that both extensive text preprocessing prior to sentiment analysis tasks and improved algorithms that take noisy text into account seem to be reasonable in order to cope with texts published by users on social media platforms.

In further research work we shall address: (i) measure and evaluate implications of noisy texts on more algorithms in order to find a set of useful preprocessing techniques that improve sentiment analysis; (ii) increase sample size and expand sample to other areas than brands in consumer electronics in order to retrieve more generalizable results; (iii) experiment and develop machine learning models that take noisy text into consideration.

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

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems (TOIS), 26*(3), 12–34.

Airoldi, E., Bai, X., & Padman, R. (2006). Markov blankets and meta-heuristics search: sentiment extraction from unstructured texts. In B. Mobasher, O. Nasraoui, B. Liu, & B. Masand (Eds.), *Lecture notes in computer science advances in web mining and web usage analysis* (pp. 167–187). Berlin, Heidelberg: Springer.

Balahur, A. (2013). Sentiment analysis in social media texts. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 120–128).

Balahur, A., Steinberger, R., van der Goot, E., Pouliquen, B., & Kabadjov, M. (2009). Opinion mining on newspaper quotations. In R. Baeza-Yates & P. Boldi (Eds.). *WI-IAT '09, Proceedings of the 2009 IEEE/WIC/ACM international joint conference on web intelligence and intelligent agent technology /// IEEE/WIC/ACM international joint conferences on web intelligence and intelligent agent technologies, 2009. WI-IAT '09; 15–18 September 2009, Milano, Italy; proceedings* (Vol. 03, pp. 523–526). Piscataway, NJ: IEEE.

Balahur, A., & Turchi, M. (2013). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*. http://dx.doi.org/10.1016/j.bbr.2011.03.031.

Banea, C., Mihalcea, R., & Wiebe, J. M. (2010). Multilingual subjectivity: Are more languages better? In *Proceedings of the 23rd international conference on computational linguistics* (pp. 28–36).

Bergsma, S., McNamee, P., Bagdouri, M., Fink, C., & Wilson, T. (2012). Language identification for creating language-specific Twitter collections. In *Proceedings of the second workshop on language in social media* (pp. 65–74).

Bollegala, D., Weir, D. J., & Carroll, J. (2011). Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *Proceedings of the 49th annual meeting of the association for computational linguistics* (pp. 132–141).

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stockmarket. *Journal of Computational Science, 2*(1), 1–8.

Boyd-Graber, J., & Resnik, P. (2010). Holistic sentiment analysis across languages: Multilingual supervised latent dirichlet allocation. In *Proceedings of the 2010 conference on empirical methods in natural language processing. EMNLP-2010* (pp. 45–55). Association for Computational Linguistics.

Brody, S., & Diakopoulos, N. (2011). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 562–570).

Caumanns, J. (1999). *A fast and simple stemming algorithm for german words*.

Chaovalit, P., & Zhou, L. (2005). Movie review mining: A comparison between supervised and unsupervised classification approaches. In *Proceedings of the 38th annual Hawaii international conference on system sciences* (pp. 112–121).

Choi, Y., & Cardie, C. (2010). Hierarchical sequential learning for extracting opinions and their attributes. In *Proceedings of the ACL 2010 conference short papers* (pp. 269–274).

Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 355–362).

Clark, A. (2003). Pre-processing very noisy text. In *Proceedings of workshop on shallow processing of large corpora* (pp. 12–22).

Cui, H., Mittal, V., & Datar, M. (2006). Comparative experiments on sentiment classification for online product reviews. In *Proceedings of AAAI-2006* (pp. 1265–1270).

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *ACL'02, Proceedings of the 40th anniversary meeting of the association for computational linguistics*.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced sentiment learning using Twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics. Posters* (pp. 241–249).

Derczynski, L., Maynard, D., Aswani, N., & Bontcheva, K. (2013). Microblog-genre noise and impact on semantic annotation accuracy. In *24th ACM conference on hypertext and social media*.

Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the international conference on recent advances in natural language processing* (pp. 198–206).

Derczynski, L., Yang, B., & Jensen, C. S. (2013). Towards context-aware search and analysis on social media data. In *Proceedings of the 16th international conference on extending database technology* (pp. 137–142).

Dey, L., & Haque, S. M. (2009). Opinion mining from noisy text data. *International Journal on Document Analysis and Recognition (IJDAR), 12*(3), 205–226.
Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *International conference on web search & data mining. Palo Alto, California, February 11–12, 2008*. New York, NY: ACM.
Gotoh, Y., & Renals, S. (2000). Sentence boundary detection in broadcast speech transcripts. In *Automatic speech REcognition: Challenges for the new millennium* (pp. 228–235).
Güngör, T. (2010). Part-of-speech tagging. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of natural language processing* (2nd ed., pp. 205–235). Boca Raton, FL: Chapman & Hall/CRC.
Guo, H., Zhu, H., Guo, Z., Zahng, X., & Su, Z. (2009). Product feature categorization with multilevel latent semantic association. In *Proceedings of the 18th ACM conference on information and knowledge management* (pp. 1087–1096).
Guo, H., Zhu, H., Guo, Z., & Su, Z. (2011). Domain customization for aspect-oriented opinion analysis with multi-level latent sentiment clues. In *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 2493–2496).
Guozheng, Z., Faming, Z., Fang, W., & Jian, L. (2008). Knowledge creation in marketing based on data mining. In *International conference on intelligent computation technology and automation* (Vol. 1, pp. 782–786).
Hatzivassiloglou, V., & Wiebe, J. M. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In: *COLING '00, Proceedings of the 18th conference on computational linguistics* (Vol. 1, pp. 299–305).
Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42*, 177–196.
Hu, M., & Liu, B. (2004a). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 168–177).
Hu, M., & Liu, B. (2004b). Mining opinion features in customer reviews. In *Proceedings of AAAI* (pp. 755–760).
Jakob, N., & Gurevych, I. (2010). Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1035–1045).
Jin, W., Ho, H. H., & Srihari, R. K. (2009). OpinionMiner: A novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1195–1204).
Jindal, N., & Liu, B. (2006b). Mining comparative sentences and relations. *AAAI'06, Proceedings of the 21st national conference on artificial intelligence* (Vol. 2, pp. 1331–1336). AAAI Press. <http://dl.acm.org/citation.cfm?id=1597348.1597400>.
Jindal, N., & Liu, B. (2006a). Identifying comparative sentences in text documents. In S. Dumas (Ed.), *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 244–251). New York, NY, USA: Association for Computing Machinery.
Kaiser, C. (2009). Opinion mining im Web 2.0 – Konzept und Fallbeispiel. *HMD – Praxis der Wirtschaftsinformatik, 46*(268), 90–99.
Kemighan, M. D., Church, K. W., & Gale, W. A. (1990). A spelling correction program based on a noisy channel model. In *Proceedings of the 13th conference on computational linguistics* (Vol. 2, pp. 205–210).
Kessler, J. S., & Nicolov, N. (2009). Targeting sentiment expressions through supervised ranking of linguistic configurations. In *Proceedings of the third international AAAI conference on weblogs and social media* (pp. 90–97).
Kim, S.-M., & Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of 20th international conference on computational linguistics* (pp. 1367–1373). Geneva, Switzerland.
Kiss, T., & Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics, 32*(4), 485–525.
Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of the fifth international AAAI conference on weblogs and social media* (pp. 538–541).
Leshed, G., & Kaye, J. (2006). Understanding how bloggers feel: Recognizing affect in blog posts. In *Conference on human factors in computing systems* (pp. 1019–1024). New York, NY, USA.
Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.-J., Zhang, S., & Yu, H., (2010). Structure-aware review mining and summarization. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 653–661).
Liu, B. (2008). *Web data mining: Exploring hyperlinks, contents, and usage data (Corr. 2. print). Data-centric systems and applications*. Berlin: Springer. <http://www.gbv.de/dms/weimar/toc/584528469_toc.pdf/http://www.zentralblatt-math.org/zmath/en/search/?an=1138.68322/http://www.gbv.de/dms/bowker/toc/9783540378815.pdf>.
Liu, B. (2012). *Sentiment analysis and opinion mining*. San Rafael: Morgan & Claypool.
Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the Web. In: *WWW '05, Proceedings of the 14th international conference on World Wide Web* (pp. 342–351). New York, NY, USA.
Maynard, D., Bontcheva, K., & Rout, D. (2012). Challenges in developing opinion mining tools for social media. In *Proceedings of @NLP can u tag #user_generated_content?! Workshop at LREC 2012*.
Maynard, D., Dupplaw, D., & Hare, J. (2013). Multimodal sentiment analysis of social media. In *BCS SGAI workshop on social media analysis*. <http://eprints.soton.ac.uk/360546/>.
McCandless, M., Hatcher, E., & Gospodnetić, O. (2010). *Lucene in action* (2nd ed.). Greenwich: Manning.
Mikheev, A. (2002). Periods, capitalized words, etc.. *Computational Linguistics, 28*(3), 289–318.
Mishne, G., & Glance, N. S. (2006). Predicting movie sales from blogger sentiment. In *Proceedings of the 21st national conference on artificial intelligence. Boston* (pp. 11–14). Massachusetts: AAAI Press.
Moghaddam, S., & Ester, M. (2010). Opinion digger: An unsupervised opinion miner from unstructured product reviews. In *Proceedings of the 19th ACM international conference on information and knowledge management* (pp. 1825–1828).
Moghaddam, S., & Ester, M. (2012). On the design of LDA models for aspect-based opinion mining. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 803–812).
Nakagawa, T., Inui, K., & Kurohashi, S. (2010). Dependency tree-based sentiment classification using CRFs with hidden variables. In *Human language technologies: The 2010 annual conference of the North American chapter of the ACL* (pp. 786–794).
Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the seventh conference on international language resources and evaluation (LREC)* (pp. 1320–1326). Valletta, Malta.
Pan, S. J., Ni, X., Sun, J.-T., Yang, Q., & Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World Wide Web* (pp. 751–760).
Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135.
Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V., & Holzinger, A. (2013). Opinion mining on the Web 2.0 – Characteristics of user generated content and their impacts. In *Lecture notes in computer science: Human–computer interaction and knowledge discovery in complex, unstructured, big data*. In A. Holzinger & G. Pasi (Eds.). *Lecture notes in computer science* (Vol. 7947, pp. 35–46). Berlin, Heidelberg: Springer.
Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Winkler, S. M., Schaller, S., et al (2012). On text preprocessing for opinion mining outside of laboratory environments. In R. Huang, A. Ghorbani, G. Pasi, T. Yamaguchi, N. Yen, & & B. Jin (Eds.), *Lecture notes in computer science. Active media technology* (pp. 618–629). Berlin, Heidelberg: Springer.
Popescu, A.-M., & Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 339–346).
Porter, M. F. (2001). Snowball: A Language for Stemming Algorithms. <http://snowball.tartarus.org/texts/introduction.html>.
Rapid-i. RapidMiner. <http://rapid-i.com/content/view/181/190/lang,en/>.
Sayeed, A. B. (2011). *A distributional and syntactic approach to fine-grained opinion mining* (Dissertation). University of Maryland.
Steinberger, J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Lenkova, P., et al (2012). Creating sentiment dictionaries via triangulation. *Decision Support Systems, 53*(4), 689–694.

Steinberger, J., Lenkova, P., Kabadjov, M., Steinberger, R., & Goot van der, Erik (2011). Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In *Proceedings of the 8th international conference recent advances in natural language processing* (pp. 770–775).

The Apache Software Foundation. *Apache OpenNLP developer documentation: Written and maintained by the apache OpenNLP development community*. <http://opennlp.apache.org/documentation/1.5.3/manual/opennlp.html>.

The Apache Software Foundation. *LUCENE.net search engine library*. <http://lucenenet.apache.org/>.

Turney, P. D. (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. M., & Mihalcea, R. (2006). Word sense and subjectivity. In: *ACL-44, Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics* (pp. 1065–1072).

Wilson, T., Wiebe, J. M., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics, 35*(3), 399–433.

Wong, T.-L., Bing, L., & Lam, W. (2011). Normalizing web product attributes and discovering domain ontology with minimal effort. In *Proceedings of the fourth ACM international conference on web search and data mining* (pp. 805–814).

Yi, L., & Liu, B. (2003). Web page cleaning for web mining through feature weighting. In *IJCAI'03, Proceedings of the 18th international joint conference on artificial intelligence* (pp. 43–48). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc..

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). *Combining lexicon-based and learning-based methods for Twitter sentiment analysis: Technical report HPL-2011-89*.

Zhang, T. (2010). Fundamental statistical techniques. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of natural language processing* (2nd ed., pp. 189–204). Boca Raton, FL: Chapman & Hall/CRC.

Zhang, W., Yu, C., & Meng, W. (2007). Opinion retrieval from blogs. In *Proceedings of the sixteenth ACM conference on conference on information and knowledge management. Lisboa, Portugal, November 6–10, 2007* (pp. 831–840). New York, NY: ACM.