Noname manuscript No.
(will be inserted by the editor)

A comparative empirical study on social media sentiment analysis over various genres and languages

Viktor Hangya · Richárd Farkas

Received: date / Accepted: date

Abstract People express their opinions about things like products, celebrities and services using social media channels. The analysis of these textual contents for sentiments is a gold mine for marketing experts as well as for research in humanities, thus automatic sentiment analysis is a popular area of applied artificial intelligence. The chief objective of this paper is to investigate automatic sentiment analysis on social media contents over various text sources and languages. The comparative findings of the investigation may give useful insights to artificial intelligence researchers who develop sentiment analyzers for a new textual source. To achieve this, we describe supervised machine learning based systems which perform sentiment analysis and we comparatively evaluate them on seven publicly available English and Hungarian databases, which contain text documents taken from Twitter and product review sites. We discuss the differences among these text genres and languages in terms of document- and target-level sentiment analysis.

Keywords Natural language processing · Sentiment analysis · Data mining

1 Introduction

Recently, the popularity of social media has increased. People post messages on a variety of topics, like products and political issues and a large amount of user generated data is created in textual form. Several applications have been developed for exploiting the knowledge and information present in user generated content, for example in predicting the results of elections (Sang and Bos, 2012), monitoring brands (Jansen et al, 2009) and disaster management (Varga et al, 2013). In this study, we focus on sentiment analysis (SA) whose task is to assign polarity labels (positive, negative and neutral) to textual elements.

V. Hangya
Department of Computer Algorithms and Artificial Intelligence, University of Szeged
Árpád tér 2., 6720 Szeged, Hungary
E-mail: hangyav@inf.u-szeged.hu

R. Farkas
E-mail: rfarkas@inf.u-szeged.hu

Sentiment analysis can be applied at different levels depending on the depth of information which we would like to extract from the texts. In our paper, we compare two types of SA tasks, one where we seek to identify the global sentiments of people and one where we are only interested in opinions which are related to a given target.

On social media sites like forums, Twitter or Facebook, people post status messages about their sentiments, life events, and so on. In the case of the so called **document level** SA, we focus on three kinds of documents (Liu, 2012). The task at this level is to decide whether a given document contains an overall positive or negative sentiment. Consider the following example:

Um I just learned that Sunday is NATIONAL CHOCOLATE DAY. I'm totally taking advantage of that. (1)

The above text expresses the idea that the author has a positive sentiment because of the forthcoming *chocolate day*. To detect this sentiment, all the sentences in the document have to be analyzed because it expresses the positive sentiment as a whole. To detect the polarity of the sentiments, features are extracted from the texts that are signs for subjective opinions and intense emotional states.

Target level SA performs a fine-grained opinion extraction (Jiang et al, 2011). In this case the output of the system is a combination of a polarity label and a referred target. The target can be an entity (person, product, service) or some aspect (battery life, quality of a service) of a given entity. In some cases this level is called entity or aspect level SA depending on the type of the target. In the following examples for both types of targets can be seen:

I do agree that money can't buy happiness. But somehow, it's more comfortable to sit and cry in a ***BMW*** *than on a bicycle!* (2)

The ***menu*** *is limited but almost all of the* ***dishes*** *are excellent.* (3)

In the first example the target entity is *BMW*, but the negative sentiment is not related to it. Because there is no sentiment towards *BMW*, the polarity label for this example is neutral. In the second example the target aspects are the *menu* and the *dishes*, which are aspects of a restaurant. The sentence contains sentiments related to both aspects, one with negative polarity and one with positive. To handle this task, first the textual parts have to be detected which are related to a given target and their polarities can be decided only in the second step. Furthermore, in some cases like the first example, there is no sentiment towards the given target. To overcome these problems, we exploited the syntactic structure of the sentences by using dependency and constituency parsers to locate text parts which are related to the given target.

The chief contribution of this study is the comparative evaluation of task-specific techniques and their performance on various text genres and languages. We created supervised machine learning based systems for document and target-level sentiment analysis. We introduce various techniques for target-level SA and empirically investigate the added value of these special techniques over the document level methods. We carried out experiments using English databases containing Twitter messages and forum posts respectively as well as databases containing

Hungarian texts. We analyze our sentiment analysis systems on these genres and typologically different languages in a comparative way.

Opposite to product reviews, tweets are created almost in real-time so their form is less standard and they contain many more spelling errors, slang and other out of vocabulary words. Syntactic parsers are trained on standard texts so their accuracy on tweets is lower (Foster et al, 2011) as well. We propose a distance-based reweighing method to determine expression-target relatedness for tweets.

In Hungarian texts other difficulties have to be addressed. Hungarian is a free word order morphologically rich language. It has several word forms, which may mean that some words are not seen in the training phase. A word's syntactic role is defined by its morphology, unlike English, where word order is determinative.

The paper is organized as follows. The next section provides a discussion of the related work. In Section 3, the proposed techniques will be described for both document and target-level SA. Then the databases used for the comparative evaluation are introduced in Section 4. In Section 5, we focus on the results achieved and we discuss them in more detail in Section 6. Lastly, in Section 7, we draw some final conclusions.

2 Related Work

Owing to its direct marketing applications, sentiment analysis (Liu, 2012) has become an active area of research. O'Connor and Balasubramanyan (2010) showed that public opinion correlates with sentiments extracted from texts. Sentiment analysis using social media can capture large scale trends using the large amount of data generated by people. In (Jansen et al, 2009) consumer opinions from microblogs concerning various brands were investigated. It was shown that 19% of microblog messages contain the mention of a brand and 20% of these contain sentiments related to the brand. Monitoring these sentiments allows companies to gain insights into the positive and negative aspects of their products. Furthermore, analyzing microblogs permits political parties to manage their campaign better. For example, Sang and Bos (2012) used Twitter messages to predict the outcome of the Dutch election. It was shown that the results became nearly as good as traditionally obtained opinion polls. There are numerous publications about this topic and its applications; see (Liu, 2012; Feldman, 2013) for reviews which summarizes the general sentiment analysis problems and methods, and (Ravi and Ravi, 2015) for a more recent survey which also deals with cross-lingual sentiment analysis. Because tweets are forming a specific text genre there are many papers about the difficulties and method of SA in Twitter (Martínez-Cámara et al, 2012; Montejo-Ráez et al, 2014). In our paper, we comparatively evaluated task-specific techniques and their performance on various text genres and languages.

In the past few years shared tasks were organized to promote research in SA for social media. The goal of *SemEval-2014 Task 9 – Sentiment Analysis in Twitter* (Rosenthal et al, 2014) was to classify a given Twitter message into positive, negative or neutral classes. In the contribution of (Hangya et al, 2013) it was shown that Twitter specific normalization of texts like URL and emoticon unification is a significant step before classification. Most of the participating systems were based on supervised machine learning techniques. The features used included word-based, syntactic and Twitter specific ones such as abbreviations and emoticons.

The systems heavily relied on word polarity lexicons like MPQA (Wilson et al, 2005), SentiWordNet (Baccianella et al, 2010) or lexicons designed specifically for this task (Zhu et al, 2014).

Often when classifying whole documents into polarity classes, counting words and phrases is not enough. In many cases the proposed system has to understand the given document so that it can correctly determine its polarity. For example, if a sentence contains only positive words, but they are negated or ironic, a simple n-gram based system would classify it as positive although it could be negative. Reyes and Rosso (2013) proposed a method for automatic verbal irony detection. It is based on three conceptual layers: signatures, emotional scenarios and unexpectedness. They showed that detecting irony given a single sentence is quite a hard task even for human annotators. Negations can also invert the polarity of a given sentence (Wiegand et al, 2010), hence most of the systems employ simple rules for handling the effects of negation in texts. Vilares et al (2015a) used syntax-based rules to deal with negations and word intensifiers. Instead of inverting the polarity of negated words they altered its polarity value based on the word which negates it.

By analyzing texts related to a given entity we can gain deeper insights into its positive and negative aspects. The focus of *RepLab-2013 – An evaluation campaign for Online Reputation Management Systems* (Amigó et al, 2013) shared task was on target level SA. The participants' task was to perform online reputation monitoring of a particular entity on Twitter messages. One subtask was to detect the polarity of sentiments in messages related to a given entity. The best performing systems tried to capture the contexts which are related to the given entity. Cossu et al (2013) used Continuous Context Models which tend to capture and model the positional and lexical dependencies existing between a given word and its context. In the system which achieved the best results (Hangya and Farkas, 2013), besides various features, distance weighting was used to model the context of a given entity. Jiang et al (2011) also experimented with Twitter messages which were related to celebrities, companies and products. The proposed SVM-based classifier incorporated target-dependent features which relied on the syntactic parse tree of the sentences. Since tweets are usually short texts, related tweets (retweets, reply to or replied by tweets) were also taken into consideration in the classification phase. Also *SemEval-2014 Task 4 – Aspect Based Sentiment Analysis* focused on target-level SA (Pontiki et al, 2014). The goal of this task was to identify the aspects of given target entities(Poria et al, 2014; Li et al, 2015) and the sentiment expressed for each aspect. The data provided by the organizers consists of restaurant and laptop reviews. In subtask 2 (detecting the polarity of sentiments related to a given aspect), the best performing systems (Wagner et al, 2014; Kiritchenko et al, 2014) were based on SVM classifiers. The features used were the following: n-grams; target-dependent features (using parse trees, window of n words surrounding the aspect term); polarity lexicon based features. It was shown that the most useful features were those derived from the lexicons. Other systems (Hangya et al, 2014) exploited constituency parse trees by selecting constituents which are related to the aspect in question.

Other important 'deep' information for SA is the relation between text parts. Lazaridou et al (2013) proposed a joint model for unsupervised induction of sentiment, aspect and discourse information. They showed that the performance can be improved by incorporating latent discourse relations (but, and, when, etc.) in

the model. Socher et al (2013) tried to capture the compositional effects of the sentences with Recursive Neural Tensor Networks. It was shown that the proposed method can capture the negation of different sentiments and can improve the accuracy of polarity detection.

3 Document and Target Level SA Systems

In the following, we present our systems for both document and target level SA. Our systems, like the state-of-the-art systems, are based on supervised machine learning techniques. We used a maximum entropy classifier with default parameter values taken from the MALLET toolkit (McCallum, 2002). We will present our results achieved by our systems in Section 5 and discuss them in Section 6.

3.1 Document level system

The aim of document level SA is to decide the polarity of sentiments in a given document globally. More formally, given a document set $\mathcal{D}$, for each document $d \in \mathcal{D}$ we have to assign a label $l \in \mathcal{L}$, where $\mathcal{L}$ is the set of polarity labels (usually positive, negative and neutral). Besides the pure text document, there is no external information given like the entity or aspects which the document is related to, so the whole text has to be analyzed to solve this problem.

3.1.1 Preprocessing

Before extracting features from the texts, we applied the following preprocessing steps:

- In order to eliminate the multiple forms of a single word, we converted them into lowercase form, except those which are fully uppercased. In the case of the Twitter corpora we also stemmed the words with the Porter stemming algorithm.
- We replaced the @ Twitter-specific tag and each URL with the *[USER]* and *[URL]* notations, respectively. Next, in the case of a hash tag we deleted the hash mark from it; for example we converted *#funny* to *funny*. This way, we did not distinguish Twitter specific tags from other words.
- Emoticons are used frequently to express sentiments. For this reason we grouped them into positive and negative emoticon classes. We treated *:), :-),:), :D, =), ;), ;), (:* and *:(, :-(, : (,):,) :* as positive and negative, respectively.
- Although the numbers can hold polarity information in some cases, we have to understand the meaning of the number in the given context to exploit it. Without deeper semantic analysis, keeping the exact value of numbers introduces data-sparsity problems hence we decided to convert them to the *[NUMBER]* form.
- We removed the unnecessary characters `'"#$%&()*+,./:;<=>\^{}~`.
- In the case of words that contained character repetitions – more precisely those that contained the same character at least three times in a row –, we reduced the length of this sequence to three. For instance, in the case of the word

yeeeeahhhhhhh we got the form *yeeeahhh*. This way, we unified these character repetitions, but we did not lose this extra information.

3.1.2 Feature set

In our supervised settings we used **n-grams** (unigram and bigram) as well as special features which characterize the polarity of the documents. One such feature is the polarity of each word in a message. To determine the **polarity of a word**, we used sentiment lexicons.

In the case of English texts, SentiWordNet was used for this purpose (Baccianella et al, 2010). In this resource, synsets – i.e. sets of word forms sharing some common meaning – are assigned positivity, negativity and objectivity scores lying in the [0, 1] interval. These scores can be interpreted as the probability of seeing some representatives of the synsets with a positive, negative and neutral meaning, respectively. However, it is not unequivocal to determine automatically which particular synset a given word belongs to in the case of its context. Consider the word form *great* for instance, which might have multiple, entirely different sentiment connotations in different contexts, e.g. in expressions such as *"great food"* and *"great crisis"*. We determined the most likely synset a particular word form belonged to based on its contexts by selecting the synset, the members of which were the most appropriate for the lexical substitution of the target word. The extent of the appropriateness of a word being a substitute for another word was measured relying on Google's N-Gram Corpus, using the indexing framework described in (Ceylan and Mihalcea, 2011). We looked up the frequencies of the n-grams that we derived from the context by replacing the target words with its synonyms (*great*) from various synsets, e.g. *good* versus *big*. We counted the frequency of the phrases *food is good* and *food is big* in a huge set of in-domain documents (Ceylan and Mihalcea, 2011). Then we chose the meaning (synset) which had the highest probability values, which was *good* in this case.

In the case of the Hungarian corpora we used a simple sentiment lexicon, which contains a set of words with their polarity values. The lexicon was constructed in-house by a linguist expert and it contains 3322 words.

After we had assigned a polarity value to each word in a text, we created two new features for the machine learning algorithm, which were the number of positive and negative words in the given document. We treated a word as positive or negative if the related positive or negative value was greater than 0.2.

We also tried to group **acronyms** according to their polarity. For this purpose, we made use of an acronym lexicon[1]. For each acronym we used the polarity of each word in the acronym's description and we determined the polarity of the acronym by calculating the rate of positive and negative words in the description. This way, we created two new features that are the number of positive and negative acronyms in a given message.

Our hypothesis was that people like to use **character repetitions** in their words to express their happiness or sadness. Besides normalizing these tokens, we created a new feature as well which represents the number of these kinds of words in a tweet.

[1] `www.internetslang.com`

Beyond character repetitions, people like to write words or a part of the text in **uppercase** in order to call the reader's attention to it. Because of this we created yet another feature which is the number of uppercase words in the given text.

Since **negations** are quite frequent in user reviews and have the tendency to flip polarities, we took special care of negation expressions. We collected a set of negation expressions like *not* and *don't*, and a set of delimiters like *and* and *or*. We think the scope of a negation starts when we detect a negation word in the sentence and it lasts until the next delimiter. If an n-gram was in a negation scope, we added a *NOT* prefix to that feature. We did not invert the polarity of a word if it is negated. Our earlier experiments showed that this technique does not improve the results. The reason is that if a word with positive or negative polarity is negated, its polarity does not necessarily get inverted; for instance *not excellent* does not necessarily means that it is awful. Vilares et al (2015a) created a syntax-based negation detector which can detect negations more precisely in case we have accurate dependency trees. We used this simpler method in order to create a robust system on various text genres.

Besides the above mentioned supervised steps we used Latent Dirichlet Allocation (LDA) (Blei et al, 2003) for topic modeling. The goal of **topic modeling** is to discover abstract topics that occur in a collection of documents. From the results of LDA, we get the topic distribution for each document. We used the three most probable topic IDs as additional features.

3.2 Target level system

In the case of target level SA, opinions towards a given target (entity or its aspect) are investigated. A set of documents $\mathcal{D}$ and a set of targets $\mathcal{T}$ are given. For each $(d, t) \in \mathcal{D} \times \mathcal{T}$ document-target pair, a polarity label $l \in \mathcal{L}$ has to be chosen. A document can contain more than one potential target, such as when two entities or aspects are compared to each other. Each of them can be the target of a sentiment expression. We will assume here that the target mentions are given in each sentence and our task is to decide its polarity.

Next, we will introduce the techniques that were used additionally with those described in Section 3.1. First, we used the **target names** as feature. This way, we can incorporate apriori knowledge, about whether people usually like or dislike the given target. Then, we devised two methods to recognize parts of the text that are relevant for the target mention.

3.2.1 Distance-weighted bag-of-words features

The relevance of a token in a sentence can be characterized by the distance between the token in question and the mention of the target. The closer a token is to the target, the more likely that the given token is somehow related to the target. For example, consider the following tweet where the positive sentiment is much closer to the target *Metallica* than the negative one:

*The **Metallica** concert was awesome although I had the worst hangover next day!* (4)

For this we used weighted feature vectors, and weighted each n-gram feature by its distance in tokens from the mention of the given aspect (Hangya and Farkas, 2013):

$$w(i) = \frac{1}{e^{\frac{1}{n}|i-j|}}, \tag{1}$$

where n is the length of the sentence and the values i, j are the positions of the actual word and the given target.

3.2.2 Syntax-based features

Distance-weighting is a simple method for approximating the relevance of a text fragment from the target mention point of view. To create a more precise model we applied deep linguistic analysis and employed dependency and constituency syntactic parsers. A feature which can indicate the polarity of an opinion is the **polarity of words' modifiers**. We defined a feature template for tokens whose syntactic head is present in our positive or negative lexicon.

Since **adjectives** are good indicators of opinion polarity, we add those to our feature set that are in close proximity with the given target term. We define the proximity between an adjective and an aspect term as the length of the non-directional path between them in the dependency tree. We gather adjectives in proximity less than 6. An example of dependency can be seen in Figure 1.

We attempted to identify clauses which refer to the target. In a sentence we can express our opinions about more than one target, so it is important not to use clauses containing opinions about other targets. We developed a simple rule-based method for selecting the appropriate **subtree** (Hangya et al, 2014) from the constituent parse of the sentence in question (see Figure 2). In this method, the root of this subtree is the leaf which contains the given target initially. In subsequent steps, the subtree containing the target in its yield gets expanded until the following conditions are met:

- The yield of the subtree consists of at least five tokens.
- The yield of the subtree does not contain any other target besides the five-token window frame relative to the target in question.
- The current root node of the subtree is either the non-terminal symbol `PP` or `S` in English and `CP` or `NP` in Hungarian.

Relying on these identified subtrees, we introduced novel features. We created additional n-gram features from the yield of the subtree. In the case of English texts, we determined the **polarity of this subtree** with a method proposed by (Socher et al, 2013) and used it as a feature.

For English texts we used the Bohnet dependency parser (Bohnet, 2010) and the Stanford constituency parser (Klein and Manning, 2003). For Hungarian, we used the Magyarlanc (Zsibrita et al, 2013) and Szántó and Farkas (2014) parsers for dependency and constituency, respectively.

Here we exploited that the dependency and constituency syntactic representations have different goals and strengths (Farkas and Bohnet, 2012). We propose to use two separated parsers for this purpose. In computational resource demanding scenarios, it is worth to investigate whether using just one parser and then automatically converting its output to the other representation – instead of running two parsers – is still accurate enough.

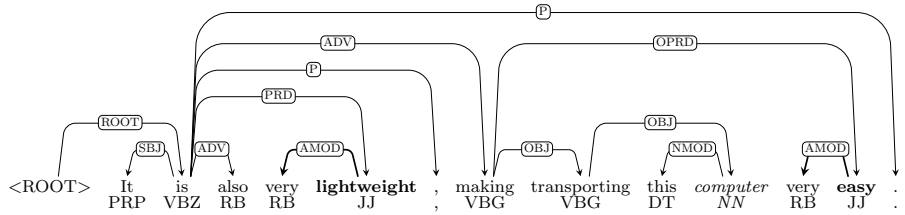

Fig. 1 Dependency parse tree (Bohnet parser), where *computer* is the target. Adjectives that are in close proximity are indicated in bold and also the relations whose head word is in the sentiment lexicon.

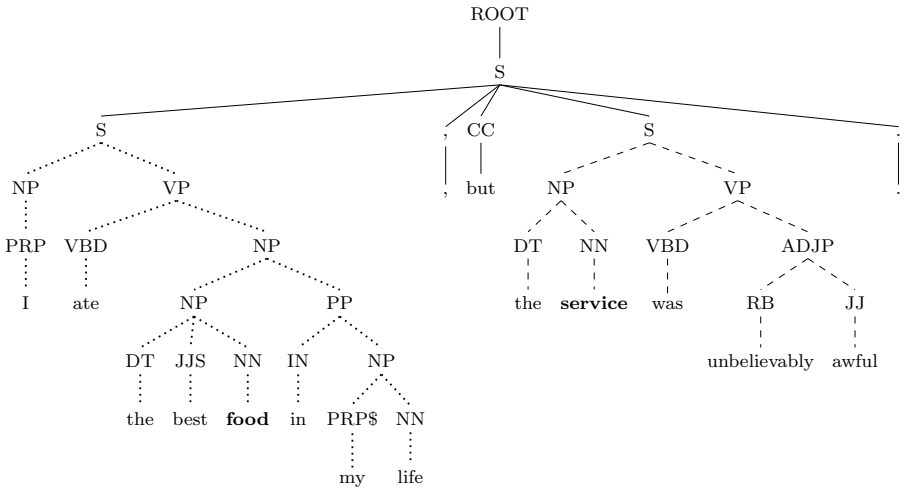

Fig. 2 Constituency parse tree (Stanford parser), with dotted and dashed subtrees for the *food* and *service* targets, respectively.

4 Databases

Next, we present the databases that were used in our experiments. For both document level and target level we used English databases consisting of Twitter messages and product reviews. We also experimented with Hungarian texts on both levels and examined their results.

The texts in the document level corpora are annotated regarding their global sentiments. In other words sentiment expressions are not referring to any target. In our experiments our goal was to analyze all the sentiments in the texts and classify them.

In case of the target level corpora texts were annotated regarding the polarity of opinions which are related to the given target. For example, if a text instance contains both positive and negative opinions but only the positive one is related to the target in question then the text is annotated with positive label. Furthermore, if a sentence contains multiple targets then there is a classification instance for all of the targets which can have different labels also. Our goal in this level was to use

methods which are aware of the target. We show that by extracting target-specific features from instances we could improve our results comparing to a system which is not aware of the target.

4.1 Document level

- **Amazon**: Product review sites are popular for both expressing opinions about products and for gathering information about positive and negative aspects before buying one. In our experiments we used 50,000 Amazon reviews which are related to DVDs and kitchen products (Jindal et al, 2008). Amazon has a 5-level rating scale, 1 being the worst and 5 being the best. Relying on these ratings we treated reviews which were given a 5 or 1 as positive or negative, respectively. We ignored all reviews between these two levels.
- **SemEval**: In 2013 the organizers of *SemEval-2013 Task 2 – Sentiment Analysis in Twitter* created a database for document level SA (Wilson et al, 2013). The corpus consists of 20,000 English Twitter messages on a range of topics. The messages were annotated with positive, negative or neutral labels depending on their global sentiments. We downloaded the training portion of it, which consisted of 10,736 messages.
- **ProdRev**: A popular Hungarian product review site is Árukereső[2]. Reviews can be found on many product categories on this site. We downloaded reviews from the PC or electronic products (TV, digital cameras, etc.) categories. The reviewers on this site have to provide pros and cons when writing a review. Here, we used these as positive and negative texts, respectively. Furthermore, we applied filtering on the texts in such a way that we only kept reviews that were one sentence long. The resulting database consisted of 6,722 documents.

4.2 Target level

- **RepLab**: For target level SA, we used a Twitter database created for the *RepLab-2013* shared task (Amigó et al, 2013). The collection comprises English and Spanish tweets on 61 entities taken from four domains: automotive, banking, universities and music. For our study we used the English training portion of the database provided, which consisted of 27,520 tweets. For each of the tweets, an entity was given which was the target of the sentiments in the text. The polarity labels could be positive, negative or neutral.
- **OpinHu**: OpinHuBank is a corpus created directly for sentiment analysis purposes (Miháltz, 2013), which consists of sentences taken from various Hungarian news, blogs and forum sites. Each sentence is at least 7 token long, has proper punctuation at its end and has at least one person entity that is the target in that sentence. The sentences were annotated for three polarity levels. The corpus contains 10,006 text instances.
- **ABSA**: The organizers of the *SemEval-2014 Task-4 – Aspect Based Sentiment Analysis* created a fine-grained corpus which contained 3,045 sentences taken from laptop reviews and 3,041 restaurant reviews (Pontiki et al, 2014). For

[2] `http://www.arukereso.hu`

each review, aspects of an entity are annotated, such as the battery life of a laptop. In this case the aspect mentions are the targets of the sentiment analysis task. For each aspect notation the polarity level is given depending on the sentiments related to the given aspect in that review. In this database, 4 polarity levels were used which were positive, negative, neutral and conflict (when both positive and negative sentiments were present).

– **JDPA**: We experimented with *The J.D. Power and Associates Sentiment Corpus* (Kessler et al, 2010), which consists of blog posts containing opinions about automobiles and digital cameras. In this corpus, sentiment expressions are annotated along with their polarities and targets. We used these annotations to create a dataset for the target level SA task. We took texts that contained at least one sentiment expression. In our setting, for each text instance, a target and a polarity level were given as earlier. We created an instance for each sentiment expression in the corpus whose target is the annotated target of the given sentiment expression and the label is its polarity level (positive or negative). This corpus can contain the same sentence several times because a sentence can contain more than one sentiment expression. We only used posts about automobiles, and this way we got 5,932 text instances.

5 Results

Now we will present the results got from using the document and target level systems. Basic statistics about the corpora used can be seen in Table 1 and 2. The average document length in characters is similar in all databases except the one from Amazon, which has more than one sentence in a document instance unlike the other corpora from the review and news genres which have only one. Furthermore, tweets are also short because of the limit in message length. In the following tables we will list the accuracy and F-scores which we obtained using 10-fold cross-validation on each corpus. We calculated per-label F_1 scores and macro-averaged F_1 score which is the unweighted average of the per-label F_1 scores. In the rest of the paper we will refer as F-score or simply as F to the macro-averaged F_1 score and we indicate the label in case of the per-label F-scores.

Table 1 Basic statistics about the corpora used. The columns show whether a corpus is annotated with target mentions, its language and genre. We have also listed the average number of words in a document and the number of labels in the corpus.

	target	language	genre	avg. doc. length	#labels
Amazon	✗	EN	review	202.31	2
SemEval	✗	EN	tweet	17.40	3
ProdRev	✗	HU	review	9.68	2
RepLab	✓	EN	tweet	14.62	3
OpinHu	✓	HU	news	26.36	3
ABSA	✓	EN	review	18.27	4
JDPA	✓	EN	review	25.81	2

Here, we created two baselines: *MFC* assigns the most frequent class (in the training set) to each document and *unigram baseline* is the maximum entropy classifier employing exclusively unigram features to assign labels to documents. In

Table 2 Label distribution and the overall number of annotated documents in each corpora which were used to our experiments.

	Positive	Negative	Neutral	Conflict	Overall
Amazon	42,713	7,287	-	-	50,000
SemEval	4,025	1,655	5,056	-	10,736
ProdRev	3,573	3,149	-	-	6,722
RepLab	16,362	3,630	7,528	-	27,520
OpinHu	882	1,629	7,495	-	10,006
ABSA	3,169	1,682	1,099	136	6,086
JDPA	3,000	2,932	-	-	5,932

Table 3 Results of three systems for the document level corpora. The accuracy (acc) scores and macro-averaged F-scores (F) were calculated using 10-fold cross-validation. Differences among systems can be seen in parentheses. The unigram baseline was compared with most frequent class (MFC) system, while the document-level was compared with the unigram baseline.

	MFC		Unigram baseline				Document-level			
	acc	F	acc		F		acc		F	
Amazon	85.43	46.70	86.63	(+1.21)	74.50	(+28.42)	87.06	(+0.41)	74.23	(-0.27)
SemEval	47.09	21.35	62.29	(+15.20)	55.71	(+34.35)	63.39	(+1.09)	56.87	(+1.16)
ProdRev	53.15	34.70	89.95	(+36.80)	89.91	(+55.20)	90.76	(+0.80)	90.73	(+0.82)

Table 4 Document-level improvements compared with the unigram baseline for the document level corpora.

	macro F	positive F	negative F	neutral F	conflict F
Amazon	-0.273	0.306	-0.852	-	-
SemEval	1.160	1.398	1.254	0.829	-
ProdRev	0.817	0.661	0.974	-	-

Tables 3 and 5, the results of these systems are listed separately for the document and the target level corpora. The differences between these two baselines are shown in parentheses. It can be seen that a simple supervised classifier with unigram features can significantly outperform the MFC system. We also compared the *document-level* system with the unigram baseline in these tables, which in addition uses techniques presented in Section 3.1, namely: preprocessing, bigrams, word and acronym polarities, character repetition and uppercased words, negations and topic modeling but no target specific techniques. With these techniques we managed to further improve the results. It can be seen that for both Twitter corpora we increased the accuracy by more than 1%, which is largely due to the effect of our preprocessing step. In the case of the other more canonical corpora, preprocessing is less effective. We achieved most of the improvements in accuracy with the ABSA corpus although the F-score decreased slightly. We ran McNemar's significance test which showed that our improvements are significant at 0.05 significance level comparing to our unigram baseline with the exception of the Amazon dataset. The reason for this is that it contains longer texts than the others and our system is not fine-tuned for longer documents. Although, the improvements are not significant our results are comparable to other's work (Vinodhini and Chandrasekaran, 2012).

We also provide a more detailed comparison between unigram baseline and document-level systems in Tables 4 and 6. It can be seen that although the macro-averaged F-score decreased in the case of Amazon, we managed to increase the F-score for the positive label. Also in the case of the ABSA database, the F-score

Table 5 Results of three systems for the target level corpora.

	MFC		Unigram baseline				Document-level			
	acc	F	acc		F		acc		F	
RepLab	59.45	24.85	71.93	(+12.47)	64.93	(+40.07)	73.16	(+1.23)	65.97	(+1.03)
OpinHu	74.91	28.55	78.82	(+3.98)	59.66	(+31.1)	79.20	(+0.31)	60.64	(+0.97)
ABSA	52.07	17.12	64.08	(+12.01)	45.72	(+28.59)	66.48	(+2.39)	45.70	(-0.01)
JDPA	50.57	33.58	73.70	(+23.13)	73.66	(+40.07)	75.23	(+1.52)	75.19	(+1.53)

Table 6 Document-level improvements compared with the unigram baseline for the target level corpora.

	macro F	positive F	negative F	neutral F	conflict F
RepLab	1.035	1.101	0.825	1.178	-
OpinHu	0.976	1.287	1.534	0.107	-
ABSA	-0.019	1.039	3.878	2.637	-7.629
JDPA	1.538	1.298	1.778	-	-

was increased for the first three labels. The significant decrease in the conflict label was caused by the low number of conflict documents present in the corpus.

The results of the target level techniques that were presented in Section 3.2 are listed in Table 7. We used these techniques besides those in the document level settings. In the table, the results of the distance-weighted bag-of-words and the syntax-based features can be seen separately as well as jointly in the so-called *hybrid* system (the feature indicating the name of the target was also used in the systems). All three systems were compared with the document-level system (Table 5).

It can be seen that in most of the cases we managed to improve the results with both techniques. The syntax-based features were more useful due to their capability to identify the clauses which are related to the given target more precisely. The exception is the RepLab database, where weighting was more useful than the parser. The reason for this is that syntactic parsers tend to perform worse on the informal Twitter messages. We achieved the best results by combining the two techniques in the hybrid system. The only exception is the RepLab Twitter corpus, where we got the best results by using just the distance weighting. Similarly to Vilares et al (2015b) who experimented with syntactic features on Spanish tweets we found that syntactic parsing just confused our system. Kong et al (2014) created a Twitter specific dependency parser which can handle the characteristics of tweets. We ran the same experiment with all of our features (hybrid) on the RepLab corpus and only replaced the Bohnet dependency parser with the Twitter specific one. We got 0.18% increase in accuracy which indicates that the Twitter specific parser performs better on tweets than those which were trained on more standard texts. On the other hand, the difference is small, i.e. the simple distance weighting method performs similarly to a syntactic parser on tweets opposed to the well-formed reviews. We have to note that the reason for the relatively small improvement might be the fact that our polarity classification system is not relying heavily on dependency trees, i.e. further improvements might be achievable if these specialties would be exploited. In Table 8, F-score differences among the hybrid and the document-level systems can be seen. The F-score of conflict label for the ABSA corpus was worse due to the low number of conflict documents; however, all the other values were better.

Table 7 Results of the target level systems (distance weighted n-grams, syntax-based features and the both of them). All three systems were compared with the document-level system (differences is parentheses).

	Distance-weighting				Syntax-based				Hybrid			
	acc		F		acc		F		acc		F	
RepLab	74.12	(+0.95)	67.68	(+1.71)	74.06	(+0.89)	67.34	(+1.37)	74.09	(+0.92)	67.36	(+1.39)
OpinHu	79.94	(+0.74)	61.55	(+0.90)	80.27	(+1.06)	61.91	(+1.26)	80.74	(+1.53)	62.71	(+2.06)
ABSA	67.23	(+0.74)	46.51	(+0.80)	68.23	(+1.74)	47.51	(+1.80)	68.40	(+1.92)	47.50	(+1.80)
JDPA	75.10	(-0.12)	75.05	(-0.14)	77.32	(+2.09)	77.27	(+2.07)	77.38	(+2.15)	77.36	(+2.16)

Table 8 Target level (hybrid) improvements compared to the document level.

	macro F	positive F	negative F	neutral F	conflict F
RepLab	1.390	0.455	0.886	2.830	-
OpinHu	2.068	0.273	4.971	0.960	-
ABSA	1.801	1.216	3.182	5.071	-2.264
JDPA	2.163	1.718	2.607	-	-

6 Discussion

In the previous section we reported quantitative results achieved by state-of-the-art document and target level systems. Now we will discuss the results and also the representative examples in more detail.

6.1 Document level

After analyzing our document level results, we can conclude that the normalization step was more important on the Twitter corpora because it contained more frequent unusual notations like various emoticons and URLs that were unified here.

Lexical knowledge is a key building block of document level SA systems as they basically aggregate the polarity scores of known expressions. Gathering lexical knowledge is non-trivial as it depends on the domain, style and genre of documents as well. In our experiments we also found that the most useful feature templates for shallow SA systems are the ones based on the sentiment lexicon. In the following example, the unigram baseline system could not learn that the words *hero* and *hopefully* are positive because of their low frequency in the training database. With the lexicon-based features, the document level system managed to correctly classify this tweet as positive.

Khader Adnan is a hero, he's the Palestinian spring and hopefully the spark of a 3rd intifada. #KhaderExists (5)

Our state-of-the-art systems achieved the best results on the Amazon and on the ProdRev datasets (see Table 3). The importance of lexical knowledge is the explanation for this. The Amazon reviews are relatively long texts and people like to redundantly express their opinions about a particular topic. Hence, although the lexicon employed is not complete SA has a good chance to recognize the polarity of the document by capturing only a low ratio of opinion expressions in the document. The case of ProdRev is different. It consists of relatively short texts

but the domain (mostly PCs) is very narrow thus the simple supervised unigram model can learn the most important domain-specific expressions.

Our inter-dataset experiments also reveal that sentiment lexicon-based features are useful for classifying positive and negative texts. Tables 4 and 6 justifies this assumption, where in the case of the corpora with more than two polarity labels, the average F-score differences across these corpora are higher for the positive and negative labels (+1.53 in average) than for the neutral label (+1.19). The F-score for the conflict label on the ABSA database decreased significantly (-7.63), which led to the decrease in the macro averaged F-score. This decrease was caused by the low occurrence of the conflict label (only 2.25%) in the corpus and the fact that these texts contain both positive and negative sentiments. A future task could be to develop features which characterize conflicting sentiments related to the same target. Although the accuracy increased in the case of the Amazon database, the F-score decreased, which was caused by the low F-score for the negative label. As was shown in Table 1, reviews in this corpus were longer and people tended to use more sentences to express their opinions. In many cases the reviewer specifies the positive aspects of a product, but concludes with a negative sentiment.

The specifications of the device are good. Also the price isn't the lowest, which indicates that this should not be the worst device ever. But I regret the purchase. (6)

Our system calculates the rate of positive and negative features and this way these kinds of reviews were classified wrongly. A solution to this problem is to analyze the sentiment of each sentence individually and use the aggregated values as the document's sentiment. Zhang et al (2009) introduced a system where they aggregated the sentence level sentiment values based on several features. They showed that sentences which are at the beginning or at the end of a document and the ones which are in first person are more important during the aggregation.

6.2 Target level

Document level features are meant to capture the global sentiment of the documents, so they cannot handle opinion about a particular target entity which is expressed locally in a document. In our target level systems, we first tried to capture and emphasize clauses which were relevant. In the following example the targets of the document are *controls*, *gauges* and *numeric counter*. The document-level system manages to decide the global polarity of the sentence, which is positive in the example, but it assigns this polarity to all the targets because the same features are extracted in each case. The target level system can differentiate by extracting different features for each target. In the example, the yield of the selected syntactic subtree for the target gauges is *to be in a good place: easy to read gauges*. Using only this clause it is clear for the classifier that the sentiments related to the target are positive. Similarly for the target numeric counter, the sentiment is clear from the yield *the numeric counter on the speedo was small*. A more complex task is to decide the polarity related to the target control because the yield of the selected subtree is the whole sentence that contains both positive and negative sentiments. In this case, our distance-weighted bag-of-words features emphasize the positive

good and *easy* words against the negative word *small*. Using these techniques we improved the accuracy by 1.63% in average.

*All the **controls** looked to be in a good place: easy to read **gauges**, although the **numeric counter** on the speedo was small.* (7)

Table 8 shows that the average differences in F-score are +1.83 for positive and negative labels, which is lower than +2.95 for the neutral label. From this, we may conclude that target level features are useful for deciding whether a sentiment in the text is related to the target in question because we were able to classify more precisely those cases that were neutral, but were classified as positive or negative by the document-level system.

6.3 Genre differences

In Table 7 we saw that in almost every case the syntax-based target level system was more accurate than the one using distance-weighted bag-of-words features. By using syntactic parsers, we managed to choose the appropriate clauses which did not contain other sentiments. The reason why the distance weighting technique results in lower accuracy scores is that it is just an approximation of relatedness in that (as it uses the whole sentence, but it emphasizes those words that are closer to the target). On the other hand, distance weighting outperformed syntactic parsing-based weighting on texts that are less formal like Twitter messages. The reasons for this are twofold. First, Tweets consists of very simple statements and non-textual elements – e.g. emoticons, hashtags – are used instead of longer linguistic expressions which results in non-grammatical texts. Second, syntactic parsers trained on canonical well-formed texts frequently fail on tweets because they use out-of-vocabulary words and ill-structured sentences. Consider the following tweet:

@princess_saraw: @Lady_Li @jt23lfc haha he knows ill clean n cook even when im dying #womens jobs ha x damn straight!!xx (8)

It contains spelling errors (*im*), slang words (*n*), user mentions (*@Lady_Li*), hashtags (#womens) and other out of vocabulary words (*haha*); furthermore, the sentences are not well separated. Our goal with the distance weighting was to give an alternative method for target awareness in situations where the performance of syntactic parsers is low. Our results support this hypothesis, namely that on the RepLab corpus, syntax-based features (with standard syntax parsers) perform worse than the distance-weighted ones. Also by using both techniques in the hybrid system we got lower results than with the distance-weighted system. In contrast, by using Twitter specific dependency parser (Kong et al, 2014) syntax-based features also improved the results which shows that these features are helpful if an accurate parser is given. On the other corpora in Table 7, the hybrid system outperformed the other two, which means that the positive effects of the individual techniques are additive. Distance-weighting causes a slight decrease in the JDPA database, which is due to the feature that indicates the name of the target. Without this feature (only distance weighting) accuracy scores improved by 0.2% compared to the document-level system. The reason for the detrimental effect of this feature is that the variance of the polarity labels for a given target is big.

6.4 Language differences

The languages of the world can placed in the so called morphological richness spectrum. At the one end there is English where grammatical roles are expressed by word order and words has a few morphological variations. On the other end there is Hungarian with free word order and rich morphology. It expresses grammatical roles by suffixes rather than word order. In our experiments, we found that the main challenge of developing a SA system to a new and typologically very different language is to find the appropriate lexical representation. In the case of Hungarian, due to its morphological richness, a word can appear in many forms, which implies that in the training phase, word forms are not seen as frequently as in English where there are not as many word forms. One solution is to use lemmas instead of word forms but it has drawbacks as well. Consider the following example and its lemmatized form:

IOS jobb mint az Android (IOS good-COMP as the Android) "IOS is better than Android"
IOS jó mint az Android (IOS good as the Android) "IOS is as good as Android" (9)

There is a negative sentiment related to the target *Android*, but if the lemmatized sentence is used, the classifier would wrongly classify it as positive because by lemmatizing *better* to *good* we lose information. Because of this, we decided to use the full word forms for n-gram features.

If accurate syntactic parsers are available for the other language target level features can be used in a similar way as for the English texts. The syntactic parsers used are effective on standard Hungarian texts, so syntax-based features can be extracted from the parse trees. The results given in Table 7 indicated that similar improvements (+1.06% acc.) can be achieved in this SA level as in English (+1.57% avg. acc.). This empirically demonstrates that our syntax-based feature templates – without any language-dependent tuning – are general enough to work well independently of the language(s) chosen.

6.5 Error analysis

We manually investigated incorrectly classified documents by our best performing system in order to reveal typical and critical sources of failures. We uniform randomly sampled 100 incorrectly classified documents from both the document-level and target-level corpora. Based on manual analysis of these documents, we recognized four main error categories along with a miscellaneous category with errors hard to put into any category. In Table 9 the error distributions can be seen separately for the document and target level corpora.

We came to the conclusion that most of the errors (overall 29%) occurred in cases where common-sense, domain-specific **background knowledge** or irony was utilized by the human author to express his/her opinion. A frequent case is when comparing something to another one. Here a background knowledge is required about the entity against which we are comparing it.

My new phone is as good as the Nokia 3210. (10)

Table 9 Proportion of error categories on document and target level corpora in the first two columns. The overall error category percentages are in the third column.

	Document level	Target level	Overall
Background knowledge:	59%	20%	29%
Aprior sentiment about the target:	-	24%	18%
Lexical:	20%	18%	19%
Syntactic:	-	17%	13%
Other:	22%	21%	21%

Without knowing that *Nokia 3210* is a very old device, deciding the true polarity of the above example is hard even for a human annotator. In another example for this category, the phrase *late for Math class* on its own expresses negativeness, but if we consider the whole tweet below we can correctly interpret its true meaning. Because our system cannot handle these cases, it wrongly classified it as negative. This error category is present mostly in the document level corpora (59%) but it is also frequent in case of the target level (20%).

The Oscars nominations are going to be announced on a Thursday, at 1.30 pm. I guess I'll have to be late for Math class... :D (11)

The second most frequent category is the category of **lexical** errors. These errors are made when the classifier cannot interpret words correctly. On the one hand, this occurs when the words with sentiment meanings in the given text are rare – or unseen – in the training data thus their polarity value is difficult to learn. On the other hand, this error can occur when a word is used in a different sense or context than usually. For instance, in the following example the word *sharp* has negative meaning in the sense that it can cut someone. But in the training dataset – consisting product reviews about laptops – the word is mostly used in the sense that the resolution of the display is sharp which is positive. Hence our system classified this example wrongly as positive instead of negative.

Once open, the leading edge is razor sharp. (12)

One target specific feature used by our system in case of the target level SA task is the name of the target. This way the classifier can learn the **aprior sentiment about the target** i.e. whether people tend to speak positively or negatively about a particular entity or aspect. We showed that this feature increased the accuracy of the system. In contrast, in cases when the sentiment in the text is not referring to the target in question (i.e. the document is neutral) but the feature of the target name is extracted from the document the apriori sentiment causes that our system classifies it into positive or negative. This phenomenon takes 19% of the manually examined errors. Note that this type of error can only occur in case of the target level and it is the most frequent category in this level.

*Can I ride with you in your **BMW**?* (13)

Another error category which can only occur in case of the target level is due to **syntactic** errors (13%). In such cases the text parts which relates to the target in question has to be detected and analyzed in order to classify correctly. For instance,

classification is hard when comparing two targets or two aspects of a target in a sentence because both positive and negative sentiments can be expressed (like in our next example). Also in many cases sentiment is expressed but not referring to the target in question thus there is not any sentiment related to the target at all. To overcome this problem we introduced the distance-weighting method and the syntax-based features but this error analysis reveals that target specific features still has to be improved.

*We were seated promptly as we had **reservations**, however after that the service was slow.* (14)

We created the category **other** for miscellaneous errors (21%) which are hard to categorize. For instance, example 15 was incorrectly classified to negative instead of positive although the word *worth* is present in the sentence and it cannot fit into any of the four main error categories. Example 16 was manually annotated by the dataset providers to neutral for the target *Boot Camp* which is questionable, we agree with our classifier which predicted positive label.

And the fried clams had just enough kick to them to make 'em worth eating. (15)

*BUT there's this application called **Boot Camp** which allows you to add another OS X like Windows.* (16)

7 Conclusions

Social media provides a big amount of user-generated texts on a wide variety of topics. The analysis of these opinionated texts is important in many areas, which explains why sentiment analysis has become a popular research area. SA can operate at multiple levels depending on the information that we would like to extract from the documents. In the case of the document-level SA, the focus is on the global sentiments expressed by people. However, when it comes to target-level SA, sentiments which are related to a given target entity or its aspects are examined.

In this paper, we introduced systems for both document-level and target-level SA and comparatively analyzed them on corpora with different genres and languages. Because of the informality of Twitter corpora first we applied a preprocessing step to unify Twitter specific notations. With the document-level SA, we extracted shallow information from the documents that may indicate the polarity of the texts. This information was used in a supervised machine learning-based system to classify documents into polarity classes. For the target level SA system, we introduced novel techniques for detecting sentiments related to a given target. We employed a syntactic parser to select clauses that are related to the target in question and used these clauses to detect sentiments. For corpora on which parsers perform less well, we introduced an alternative method to emphasize relevant clauses, by setting the importance of the bag-of-words features based on their distance from the target mention in the text.

We ran our systems on seven different corpora which consisted of texts taken from review, news and Twitter genres and from the English and Hungarian lan-

guages. We found that among the document level features those based on the sentiment lexicon were the most useful. With these features we managed to distinguish positive and negative sentiments better. Target level techniques were useful to detect whether the sentiments in a text are related to the given target or not. Out of the two techniques introduced, the syntax-based one performed better than distance-weighting, but the best results were achieved by using both feature sets. The only exception was the RepLab Twitter corpus where distance weighting resulted in the best performance. We also tried a Twitter specific dependency parser (Kong et al, 2014) in case of the RepLab corpus and the results showed that syntax-based features are helpful for tweets also if an accurate parser is given. In the case of the Hungarian corpora, systems with Hungarian specific linguistic preprocessing tools achieved similar results.

In summary we can conclude that the accuracy which can be expected from a state-of-the-art sentiment analyzer is highly dependent on the

- level of the analysis (as the target-level systems have to address a more difficult task than document-level systems which are usually based only on lexicon lookups),
- length of the documents (as redundant opinion expression can be exploited in longer documents),
- diversity of the domain targeted (as narrow domain's lexical knowledge can be captured from relatively few training examples),
- genre (as various genre use various grammatical complexity and they require special preprocessing steps),
- language of the texts (as language-specific lexical look-up strategies can be required and highly accurate syntactic parsers have to be available).

References

Amigó E, Carrillo de Albornoz J, Chugur I, Corujo A, Gonzalo J, Martín T, Meij E, de Rijke M, Spina D, Amigo E, de Albornoz JC, Martin T, de Rijke M (2013) Overview of replab 2013: Evaluating online reputation monitoring systems. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp 333–352

Baccianella S, Esuli A, Sebastiani F (2010) SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. the Journal of machine Learning research 3:993–1022

Bohnet B (2010) Top Accuracy and Fast Dependency Parsing is not a Contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China, pp 89–97

Ceylan H, Mihalcea R (2011) An Efficient Indexer for Large N-Gram Corpora. In: ACL (System Demonstrations), pp 103–108

Cossu JV, Bigot B, Bonnefoy L, Morchid M, Bost X, Senay G, Dufour R, Bouvier V, Torres-Moreno JM, El-Beze M (2013) LIA@RepLab 2013. In: Working Notes of CLEF 2013 Evaluation Labs and Workshop

Farkas R, Bohnet B (2012) Stacking of Dependency and Phrase Structure Parsers. In: Proceedings of COLING 2012, The COLING 2012 Organizing Committee, Mumbai, pp 849–866

Feldman R (2013) Techniques and applications for sentiment analysis. Communications of the ACM 56(4):82, DOI 10.1145/2436256.2436274

Foster J, Çetinoglu Ö, Wagner J, Le Roux J, Hogan S, Nivre J, Hogan D, Van Genabith J (2011) # hardtoparse: POS Tagging and Parsing the Twitterverse. In: AAAI 2011 Workshop on Analyzing Microtext, pp 20–25

Hangya V, Farkas R (2013) Filtering and Polarity Detection for Reputation Management on Tweets. In: Working Notes of CLEF 2013 Evaluation Labs and Workshop

Hangya V, Berend G, Farkas R (2013) SZTE-NLP: Sentiment Detection on Twitter Messages. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp 549–553

Hangya V, Berend G, Varga I, Farkas R (2014) SZTE-NLP: Aspect level opinion mining exploiting syntactic cues. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, pp 610–614

Jansen BJ, Zhang M, Sobel K, Chowdury A (2009) Twitter Power: Tweets as Electronic Word of Mouth. In: Journal of the American society for information science and technology, pp 2169–2188

Jiang L, Yu M, Zhou M, Liu X, Zhao T (2011) Target-dependent Twitter Sentiment Classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp 151–160

Jindal N, Liu B, Street SM (2008) Opinion Spam and Analysis. Proceedings of the 2008 International Conference on Web Search and Data Mining

Kessler JS, Eckert M, Clark L, Nicolov N (2010) The 2010 ICWSM JDPA Sentiment Corpus for the Automotive Domain. In: 4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)

Kiritchenko S, Zhu X, Cherry C, Mohammad S (2014) NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), SemEval, p 437

Klein D, Manning CD (2003) Accurate Unlexicalized Parsing. In: Proceedings of the 41st ACL, pp 423–430, DOI 10.3115/1075096.1075150

Kong L, Schneider N, Swayamdipta S, Bhatia A, Dyer C, Smith NA (2014) A Dependency Parser for Tweets. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, pp 1001–1012

Lazaridou A, Titov II, Sporleder CC (2013) A Bayesian Model for Joint Unsupervised Induction of Sentiment, Aspect and Discourse Representations. In: 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, pp 1630–1639

Li S, Zhou L, Li Y (2015) Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures. Information Processing & Management 51(1):58–67, DOI 10.1016/j.ipm.2014.08.005

Liu B (2012) Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies 5(1):1–167

Martínez-Cámara E, Martín-Valadivia MT, Urena-López LA, Montejo-Ráez AR (2012) Sentiment analysis in Twitter. Natural Language Engineering 20(01):1–28, DOI 10.1017/S1351324912000332

McCallum AK (2002) MALLET: A Machine Learning for Language Toolkit

Miháltz M (2013) OpinHuBank: szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez. In: IX. Magyar Számítógépes Nyelvészeti Konferencia, pp 343–345

Montejo-Ráez A, Martínez-Cámara E, Martín-Valdivia MT, Ureña-López LA (2014) A knowledge-based approach for polarity classification in Twitter. Journal of the Association for Information Science and Technology 65(2):414–425, DOI 10.1002/asi.22984

O'Connor B, Balasubramanyan R (2010) From tweets to polls: Linking text sentiment to public opinion time series. ICWSM

Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S (2014) Semeval-2014 task 4: Aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), SemEval '14, pp 27–35

Poria S, Cambria E, Ku LW, Gui C, Gelbukh A (2014) A Rule-Based Approach to Aspect Extraction from Product Reviews. In: Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP), Association for Computational Linguistics and Dublin City University, Dublin, pp 28–37

Ravi K, Ravi V (2015) A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems 89:14–46, DOI 10.1016/j.knosys.2015.06.015

Reyes A, Rosso P (2013) On the difficulty of automatically detecting irony: beyond a simple case of negation. Knowledge and Information Systems 40(3):595–614, DOI 10.1007/s10115-013-0652-8

Rosenthal S, Nakov P, Ritter A, Stoyanov V (2014) Semeval-2014 task 9: Sentiment analysis in twitter. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), SemEval, pp 73–80

Sang ETK, Bos J (2012) Predicting the 2011 Dutch Senate Election Results with Twitter. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp 53–60

Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp 1631–1642

Szántó Zs, Farkas R (2014) Special Techniques for Constituent Parsing of Morphologically Rich Languages. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp 135–144

Varga I, Sano M, Torisawa K, Hashimoto C, Ohtake K, Kawai T, Oh JH, De Saeger S (2013) Aid is Out There: Looking for Help from Tweets during a Large Scale Disaster. In: Proceedings of the 51st Annual Meeting of the ACL, pp 1619–1629

Vilares D, Alonso MA, Gómez-Rodriguez C (2015a) A syntactic approach for opinion mining on Spanish reviews. Natural Language Engineering 21(01):139–163, DOI 10.1017/S1351324913000181

Vilares D, Alonso MA, Gómez-Rodríguez C (2015b) On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages. Journal of the Association for Information Science and Technology 66(9):1799–1816,

DOI 10.1002/asi.23284
Vinodhini G, Chandrasekaran RM (2012) Sentiment Analysis and Opinion Mining: A Survey. International Journal 2(6)
Wagner J, Arora P, Cortes S (2014) DCU: Aspect-based polarity classification for semeval task 4. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) pp 223–229
Wiegand M, Balahur A, Roth B, Klakow D, Montoyo A (2010) A survey on the role of negation in sentiment analysis. In: Proceedings of the workshop on negation and speculation in natural language processing, pp 60–68
Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics, pp 347–354
Wilson T, Kozareva Z, Nakov P, Rosenthal S, Stoyanov V, Ritter A (2013) SemEval-2013 Task 2: Sentiment Analysis in Twitter. In: Proceedings of the International Workshop on Semantic Evaluation, SemEval '13
Zhang C, Zeng D, Li J, Wang FY, Zuo W (2009) Sentiment analysis of Chinese documents: From sentence to document level. Journal of the American Society for Information Science and Technology 60(12):2474–2487, DOI 10.1002/asi.21206
Zhu X, Kiritchenko S, Mohammad S (2014) NRC-Canada-2014: Recent Improvements in the Sentiment Analysis of Tweets. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp 443–447
Zsibrita J, Vincze V, Farkas R (2013) Magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP, pp 763–771