Sentiment Analysis in Czech Social Media Using Supervised Machine Learning

Ivan Habernal
NTIS – New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň Czech Republic
habernal@kiv.zcu.cz

Tomáš Ptáček
Department of Computer Science and Engineering, Faculty of Applied Sciences University of West Bohemia, Univerzitní 8, 306 14 Plzeň Czech Republic
tigi@kiv.zcu.cz

Josef Steinberger
NTIS – New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň Czech Republic
jstein@kiv.zcu.cz

Abstract

This article provides an in-depth research of machine learning methods for sentiment analysis of Czech social media. Whereas in English, Chinese, or Spanish this field has a long history and evaluation datasets for various domains are widely available, in case of Czech language there has not yet been any systematical research conducted. We tackle this issue and establish a common ground for further research by providing a large human-annotated Czech social media corpus. Furthermore, we evaluate state-of-the-art supervised machine learning methods for sentiment analysis. We explore different pre-processing techniques and employ various features and classifiers. Moreover, in addition to our newly created social media dataset, we also report results on other widely popular domains, such as movie and product reviews. We believe that this article will not only extend the current sentiment analysis research to another family of languages, but will also encourage competition which potentially leads to the production of high-end commercial solutions.

1 Introduction

Sentiment analysis has become a mainstream research field in the past decade. Its impact can be seen in many practical applications, ranging from analyzing product reviews (Stepanov and Riccardi, 2011) to predicting sales and stock markets using social media monitoring (Yu et al., 2013). The users' opinions are mostly extracted either on a certain polarity scale, or binary (positive, negative); various levels of granularity are also taken into account, e.g., document-level, sentence-level, or aspect-based sentiment (Hajmohammadi et al., 2012).

Most of the research in automatic sentiment analysis of social media has been performed in English and Chinese, as shown by several recent surveys, i.e., (Liu and Zhang, 2012; Tsytsarau and Palpanas, 2012). For Czech language, there have been very few attempts, although the importance of sentiment analysis of social media became apparent, i.e., during the recent presidential elections [1]. Many Czech companies also discovered a huge potential in social media marketing and started launching campaigns, contests, and even customer support on Facebook—the dominant social network of the Czech online community with approximately 3.5 million users.[2] However, one aspect still eludes many of them: automatic analysis of customer sentiment of products, services, or even a brand or a company name. In many cases, sentiment is still labeled manually, according to our information from one of the leading Czech companies for social media monitoring.

Automatic sentiment analysis in the Czech environment has not yet been thoroughly targeted by the research community. Therefore it is necessary to create a publicly available labeled dataset as well as to evaluate the current state of the art for two reasons. First, many NLP methods must deal with high flection and rich syntax when processing the Czech language. Facing these issues may lead to novel

[1] http://www.mediaguru.cz/2013/01/analyza-facebook-rozhodne-o-volbe-prezidenta/ [in Czech]

[2] http://www.czso.cz/csu/redakce.nsf/i/uzivatele_facebooku [in Czech]

Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 65–74, Atlanta, Georgia, 14 June 2013.

approaches to sentiment analysis as well. Second, freely accessible and well-documented datasets, as known from many shared NLP tasks, may stimulate competition which usually leads to the production of cutting-edge solutions.[3]

This article focuses on document-level sentiment analysis performed on three different Czech datasets using supervised machine learning. As the first dataset, we created a Facebook corpus consisting of 10,000 posts. The dataset was manually labeled by two annotators. The other two datasets come from online databases of movie and product reviews, whose sentiment labels were derived from the accompanying star ratings from users of the databases. We provide all these labeled datasets under Creative Commons BY-NC-SA licence[4] at `http://liks.fav.zcu.cz/sentiment` , together with the sources for all the presented experiments.

The rest of this article is organized as follows. Section 2 examines the related work with a focus on the Czech research and social media. Section 3 thoroughly describes the datasets and the annotation process. In section 4, we list the employed features and describe our approach to classification. Finally, section 5 contains the results with a thorough discussion.

2 Related work

There are two basic approaches to sentiment analysis: dictionary-based and machine learning-based. While dictionary-based methods usually depend on a sentiment dictionary (or a polarity lexicon) and a set of handcrafted rules (Taboada et al., 2011), machine learning-based methods require labeled training data that are later represented as features and fed into a classifier. Recent attempts have also investigated semi-supervised methods that incorporate auxiliary unlabeled data (Zhang et al., 2012).

2.1 Supervised machine learning for sentiment analysis

The key point of using machine learning for sentiment analysis lies in engineering a representative set of features. Pang et al. (2002) experimented with unigrams (presence of a certain word, frequencies of words), bigrams, part-of-speech (POS) tags, and adjectives on a Movie Review dataset. Martineau and Finin (2009) tested various weighting schemes for unigrams based on TFIDF model (Manning et al., 2008) and proposed delta weighting for a binary scenario (positive, negative). Their approach was later extended by Paltoglou and Thelwall (2010) who proposed further improvement in delta TFIDF weighting.

The focus of the current sentiment analysis research is shifting towards social media, mainly targeting Twitter (Kouloumpis et al., 2011; Pak and Paroubek, 2010) and Facebook (Go et al., 2009; Ahkter and Soria, 2010; Zhang et al., 2011; López et al., 2012). Analyzing media with very informal language benefits from involving novel features, such as emoticons (Pak and Paroubek, 2010; Montejo-Ráez et al., 2012), character n-grams (Blamey et al., 2012), POS and POS ratio (Ahkter and Soria, 2010; Kouloumpis et al., 2011), or word shape (Go et al., 2009; Agarwal et al., 2011).

In many cases, the gold data for training and testing the classifiers are created semi-automatically, as in, e.g., (Kouloumpis et al., 2011; Go et al., 2009; Pak and Paroubek, 2010). In the first step, random samples from a large dataset are drawn according to presence of emoticons (usually positive and negative) and are then filtered manually. Although large high-quality collections can be created very quickly using this approach, it makes a strong assumption that every positive or negative post must contain an emoticon.

Balahur and Tanev (2012) performed experiments with Twitter posts as part of the CLEF 2012 RepLab[5]. They classified English and Spanish tweets by a small but precise lexicon, which contained also slang, combined with a set of rules that capture the manner in which sentiment is expressed in social media.

[3] E.g., named entity recognition based on Conditional Random Fields emerged from CoNLL-2003 named entity recognition shared task.

[4] `http://creativecommons.org/licenses/by-nc-sa/3.0/`

[5] `http://www.limosine-project.eu/events/replab2012`

Since the limited space of this paper does not allow us to present detailed evaluation from the related work, we recommend an in-depth survey by Tsytsarau and Palpanas (2012) for actual results obtained from the abovementioned methods.

2.2 Sentiment analysis in Czech environment

Veselovská et al. (2012) presented an initial research on Czech sentiment analysis. They created a corpus which contains polarity categories of 410 news sentences. They used the Naive Bayes classifier and a classifier based on a lexicon generated from annotated data. The corpus is not publicly available, moreover, due to the small size of the corpus no strong conclusions can be drawn.

Steinberger et al. (2012) proposed a semi-automatic 'triangulation' approach to creating sentiment dictionaries in many languages, including Czech. They first produced high-level gold-standard sentiment dictionaries for two languages and then translated them automatically into the third language by a state-of-the-art machine translation service. Finally, the resulting sentiment dictionaries were merged by taking overlap from the two automatic translations.

A multilingual parallel news corpus annotated with opinions towards entities was presented in (Steinberger et al., 2011). Sentiment annotations were projected from one language to several others, which saved annotation time and guaranteed comparability of opinion mining evaluation results across languages. The corpus contains 1,274 news sentences where an entity (the target of the sentiment analysis) occurs. It contains 7 languages including Czech. Their research targets fundamentally different objectives from our research as they focus on news media and aspect-based sentiment analysis.

3 Datasets

3.1 Social media dataset

The initial selection of Facebook brand pages for our dataset was based on the 'top' Czech pages, according to the statistics from SocialBakers.[6] We focused on pages with a large Czech fan base and a sufficient number of Czech posts. Using Facebook Graph API and Java Language Detector[7] we acquired 10,000 random posts in the Czech language from nine different Facebook pages. The posts were then completely anonymized as we kept only their textual contents.

Sentiment analysis of posts at Facebook brand pages usually serves as a marketing feedback of user opinions about brands, services, products, or current campaigns. Thus we consider the sentiment target to be the given product, brand, etc. Typically, users' complaints hold negative sentiment, whereas joy or happiness about the brand is taken as positive. We also added another class called *bipolar* which represents both positive and negative sentiment in one post.[8] In some cases, the user's opinion, although being somehow positive, does not relate to the given page.[9] Therefore the sentiment is treated as neutral in these cases, according to our above-mentioned assumption.

The complete 10k dataset was independently annotated by two annotators. The inter-annotator agreement (Cohen's κ) between these two annotators reaches 0.66 which represents a substantial agreement level (Pustejovsky and Stubbs, 2013), therefore the task can be considered as well-defined.

The gold data were created based on the agreement of the two annotators. They disagreed in 2,216 cases. To solve these conflicts, we involved a third super-annotator to assign the final sentiment label. However, even after the third annotator's labeling, there was still no agreement for 308 labels. These cases were later solved by a fourth annotator. We discovered that most of these conflicting cases were classified as either neutral or bipolar. These posts were often difficult to label because the author used irony, sarcasm or the context or previous posts. These issues remain open.

The Facebook dataset contains of 2,587 positive, 5,174 neutral, 1,991 negative, and 248 bipolar posts, respectively. We ignore the bipolar class later in all experiments. The sentiment distribution among the

[6] http://www.socialbakers.com/facebook-pages/brands/czech-republic/

[7] http://code.google.com/p/jlangdetect/

[8] For example *"to bylo moc dobry ,fakt jsem se nadlabla :-D skoda ze uz neni v nabidce"—"It was very tasty, I really stuffed myself :-D sad it's not on the menu anymore"*.

[9] Certain campaigns ask the fans for, i.e., writing a poem—these posts are mostly positive (or funny, at least) but are irrelevant for the desired task.

source pages is shown in Figure 1. The statistics reveal negative opinions towards cell phone operators and positive opinions towards, e.g., perfumes and ZOO.

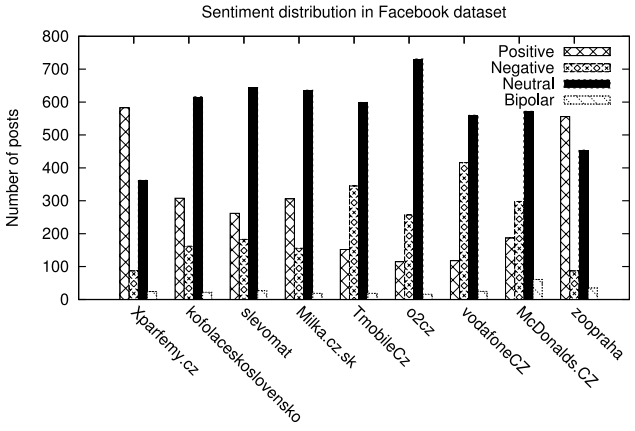

Figure 1: Social media dataset statistics

3.2 Movie review dataset

Movie reviews as a corpus for sentiment analysis has been used in research since the pioneering research conducted by Pang et al. (2002). Therefore we covered the same domain in our experiments as well. We downloaded 91,381 movie reviews from the Czech Movie Database[10] and split them into 3 categories according to their star rating (0–2 stars as negative, 3–4 stars as neutral, 5–6 stars as positive). The dataset contains of 30,897 positive, 30,768 neutral, and 29,716 negative reviews, respectively.

3.3 Product review dataset

Another very popular domain for sentiment analysis deals with product reviews (Hu and Liu, 2004). We crawled all user reviews from a large Czech e-shop Mall.cz[11] which offers a wide range of products. The product reviews are accompanied with star ratings on the scale 0–5. We took a different strategy for assigning sentiment labels. Whereas in the movie dataset the distribution of stars was rather uniform, in the product review domain the ratings were skewed towards the higher values. After a manual inspection we discovered that 4-star ratings mostly correspond to neutral opinions and 3 or less stars denote mostly negative comments. Thus we split the dataset into three categories according to this observation. The final dataset consists of 145,307 posts (102,977 positive, 31,943 neutral, and 10,387 negative).

4 Classification

4.1 Preprocessing

As pointed out by Laboreiro et al. (2010), tokenization significantly affects sentiment analysis, especially in case of social media. Although Ark-tweet-nlp tool (Gimpel et al., 2011) was developed and tested in English, it yields satisfactory results in Czech as well, according to our initial experiments on the Facebook corpus. Its significant feature is proper handling of emoticons and other special character sequences that are typical for social media. Furthermore, we remove stopwords using the stopword list from Apache Lucene project.[12]

In many NLP applications, a very popular preprocessing technique is stemming. We tested Czech light stemmer (Dolamic and Savoy, 2009) and High Precision Stemmer[13]. Another widely-used method for reducing the vocabulary size, and thus the feature space, is lemmatization. For Czech language the only currently available lemmatizer is shipped with Prague Dependency Treebank (PDT) toolkit (Hajič et al., 2006). However, we use our in-house Java HMM-based implementation using the PDT training data as we need a better control over each preprocessing step.

Part-of-speech tagging is done using our in-house Java solution that exploits Prague Dependency Treebank (PDT) data as well. However, since PDT is trained on news corpora, we doubt it is suitable for tagging social media that are written in very informal language (consult, i.e., (Gimpel et al., 2011) where similar issues were tackled in English).

Since the Facebook dataset contains a huge number of grammar mistakes and misspellings (typically *'i/y','ě/je/ie'*, and others), we incorporated phonetic transcription to International Phonetic Alphabet (IPA) in order to reduce the effect of these mistakes. We rely on eSpeak[14] implementation. An-

[10] http://www.csfd.cz/

[11] http://www.mall.cz

[12] http://lucene.apache.org/core/

[13] Publication pending; please visit http://liks.fav.zcu.cz/HPS/.

[14] http://espeak.sourceforge.net

Pipe 1	Pipe 2	Pipe 3
Tokenizing		
ArkTweetNLP		
POS tagging		
PDT		
Stem (S)		**Lemma (L)**
none (n)		PDT (p)
light (l)		
HPS (h)		
Stopwords		
remove		
Casing (C)	**Phonetic (P)**	–
keep (k)	eSpeak (e)	
lower (l)		

Table 1: The preprocessing pipes (top-down). Various combinations of methods can be denoted using the appropriate labels, e.g. "SnCk" means *1. tokenizing, 2. POS-tagging, 3. no stemming, 4. removing stopwords, and 5. no casing*, or "Lp" means *1. tokenizing, 2. POS-tagging, 3. lemmatization using PDT, and 4. removing stopwords*.

other preprocessing step might involve removing diacritics, as many Czech users type only using unaccented characters. However, posts without diacritics represent only about 8% of our datasets, thus we decided to keep diacritics unaffected.

The complete preprocessing diagram and its variants is depicted in Table 1. Overall, there are 10 possible preprocessing 'pipe' configurations.

4.2 Features

N-gram features We use presence of unigrams and bigrams as binary features. The feature space is pruned by minimum n-gram occurrence which was empirically set to 5. Note that this is the baseline feature in most of the related work.

Character n-gram features Similarly to the word n-gram features, we added character n-gram features, as proposed by, e.g., (Blamey et al., 2012). We set the minimum occurrence of a particular character n-gram to 5, in order to prune the feature space. Our feature set contains 3-grams to 6-grams.

POS-related features Direct usage of part-of-speech n-grams that would cover sentiment patterns has not shown any significant improvement in the related work. Still, POS tags provide certain characteristics of a particular post. We implemented various POS features that include, e.g., the number of nouns, verbs, and adjectives (Ahkter and Soria, 2010), the ratio of nouns to adjectives and verbs to adverbs (Kouloumpis et al., 2011), and number of negative verbs.

Emoticons We adapted the two lists of emoticons that were considered as positive and negative from (Montejo-Ráez et al., 2012). The feature captures number of occurrences of each class of emoticons within the text.

Delta TFIDF variants for binary scenarios Although simple binary word features (presence of a certain word) reach surprisingly good performance, they have been surpassed by various TFIDF-based weighting, such as Delta TFIDF (Martineau and Finin, 2009), or Delta BM25 TFIDF (Paltoglou and Thelwall, 2010). Delta-TFIDF still uses traditional TFIDF word weighting but treats positive and negative documents differently. However, all the existing related works which use this kind of features deal only with binary decisions (positive/negative), thus we filtered out neutral documents from the datasets.[15] We implemented the most promising weighting schemes from (Paltoglou and Thelwall, 2010), namely *Augmented TF*, *LogAve TF*, *BM25 TF*, *Delta Smoothed IDF*, *Delta Prob. IDF*, *Delta Smoothed Prob. IDF*, and *Delta BM25 IDF*.

4.3 Classifiers

All evaluation tests were performed using two classifiers, Maximum Entropy (MaxEnt) and Support Vector Machines (SVM). Although Naive Bayes classifier is also widely used in the related work, we did not include it as it usually performs worse than SVM or MaxEnt. We used a pure Java framework for machine learning[16] with default settings (linear kernel for SVM).

5 Results

For each combination from the preprocessing pipeline (refer to Table 1) we assembled various sets of features and employed two classifiers. In the first

[15] Opposite to leave-one-out cross validation in (Paltoglou and Thelwall, 2010), we still use 10-fold cross validation in all experiments.

[16] `http://liks.fav.zcu.cz/ml`

scenario, we classify into all three classes (positive, negative, and neutral).[17] In the second scenario, we follow a strand of related research, e.g., (Martineau and Finin, 2009; Celikyilmaz et al., 2010), that deals only with positive and negative classes. For these purposes we filtered out all the neutral documents from the datasets. Furthermore, in this scenario we evaluate only features based on weighted delta-TFIDF, as, e.g., in (Paltoglou and Thelwall, 2010). We also involved only MaxEnt classifier into the second scenario.

All tests were conducted in the 10-fold cross validation manner. We report macro F-measure, as it allows comparing classifier results on different datasets. Moreover, we do not report micro F-measure (accuracy) as it tends to prefer performance on dominant classes in highly unbalanced datasets (Manning et al., 2008), which is, e.g., the case of our Product Review dataset where most of the labels are positive.

5.1 Social media

Table 2 shows the results for the 3-class classification scenario on the Facebook dataset. The row labels denote the preprocessing configuration according to Table 1. In most cases, maximum entropy classifier significantly outperforms SVM. The combination of all features (the last column) yields the best results regardless to the preprocessing steps. The reason might be that the involved character n-gram feature captures subtle sequences which represent subjective punctuation or emoticons, that were not covered by the *emoticon* feature. On average, the best results were obtained when HPS stemmer and lowercasing or phonetic transcription were involved (lines *ShCl* and *ShPe*). This configuration significantly outperforms other preprocessing techniques for token-based features (see column *Unigr + bigr + POS + emot.*).

In the second scenario we evaluated various TFIDF weighting schemes for binary sentiment classification. The results are shown in Table 3. The three-character notation consists of term frequency, inverse document frequency, and normalization. Due to a large number of possible combinations, we report only the most successful ones, namely *Augmented*—a and *LogAve*—L term frequency, followed by *Delta Smoothed*—$\Delta(t')$, *Delta Smoothed Prob.*—$\Delta(p')$, and *Delta BM25*—$\Delta(k)$ inverse document frequency; normalization was not involved. We can see that the baseline (the first column *bnn*) is usually outperformed by any weighted TFIDF technique. Moreover, using any kind of stemming (the row entitled *various**) significantly improves the results. For the exact formulas of the delta TFIDF variants please refer to (Paltoglou and Thelwall, 2010).

We also tested the impact of TFIDF word features when added to other features from the first scenario (refer to Table 2). Column *FS1* in Table 3 displays results for a feature set with the simple binary presence-of-the-word feature (binary unigrams). In the last column *FS2* we replaced this binary feature with TFIDF weighted feature $a\Delta(t')n$. It turned out that the weighed form of word feature does not improve the performance, when compared with simple binary unigram feature. Furthermore, a set of different features (words, bigrams, POS, emoticons, character n-grams) significantly outperforms a single TFIDF weighted feature.

We also report the effect of the dataset size on the performance. We randomly sampled 10 subsets from the dataset (1k, 2k, etc.) and tested the performance; still using 10-fold cross validation. We took the most promising preprocessing configuration (*ShCl*) and MaxEnt classifier. As can be seen in Figure 2, while the dataset grows to approx 6k–7k items, the performance rises for most combinations of features. At 7k-items dataset, the performance begins to reach its limits for most combinations of features and hence adding more data does not lead to a significant improvement.

5.1.1 Upper limits of automatic sentiment analysis

To see the upper limits of the task itself, we also evaluate the annotator's judgments. Although the gold labels were chosen after a consensus of at least two people, there were many conflicting cases that must have been solved by a third or even a fourth person. Thus even the original annotators do not achieve 1.00 F-measure on the gold data.

We present 'performance' results of both annotators and of the best system as well (MaxEnt classi-

[17] We ignore the bipolar posts in the current research.

Facebook dataset, 3 classes

	Unigrams		Unigr + bigrams		Unigr + bigr + POS features		Unigr + bigr + POS + emot.		Unigr + bigr + POS + emot. + char n-grams	
	MaxEnt	SVM	MaxEnt	SVM	MaxEnt	SVM	MaxEnt	SVM	MaxEnt	SVM
SnCk	0.63	0.64	0.63	0.64	0.66	0.64	0.66	0.64	**0.69**	0.67
SnCl	0.63	0.64	0.63	0.64	0.66	0.63	0.66	0.63	**0.69**	0.68
SlCk	0.65	0.67	0.66	0.67	0.68	0.66	0.67	0.66	**0.69**	0.67
SlCl	0.65	0.67	0.65	0.67	0.68	0.66	**0.69**	0.66	**0.69**	0.67
ShCk	0.66	0.67	0.66	0.67	0.68	0.67	0.67	0.67	**0.69**	0.67
ShCl	0.66	0.66	0.66	0.67	**0.69**	0.67	**0.69**	0.67	**0.69**	0.67
SnPe	0.64	0.65	0.64	0.65	0.67	0.65	0.67	0.65	0.68	0.68
SlPe	0.65	0.67	0.65	0.67	0.68	0.67	0.67	0.66	0.68	0.67
ShPe	0.66	0.67	0.66	0.67	**0.69**	0.66	**0.69**	0.66	0.68	0.67
Lp	0.64	0.65	0.63	0.65	0.67	0.64	0.67	0.65	0.68	0.67

Table 2: Results on the Facebook dataset, classification into 3 classes. Macro F-measure, 95% confidence interval $= \pm 0.01$. Bold numbers denote the best results.

Facebook dataset, positive and negative classes only

	bnn	$a\Delta(t')n$	$a\Delta(p')n$	$a\Delta(k)n$	$L\Delta(t')n$	$L\Delta(p')n$	$L\Delta(k)n$	FS1	FS2
SnCk	0.83	0.86	0.86	0.86	0.85	0.86	0.86	**0.90**	0.89
SnCl	0.84	0.86	0.86	0.86	0.86	0.86	0.86	**0.90**	**0.90**
various*	0.85	0.88	0.88	0.88	0.88	0.88	0.88	**0.90**	**0.90**
SnPe	0.84	0.86	0.86	0.86	0.86	0.86	0.86	**0.90**	**0.90**
Lp	0.84	0.86	0.85	0.85	0.86	0.86	0.86	0.88	0.88

* same results for ShCk, ShCl, SlCl, SlPe, SlCk, and ShPe

FS1: Unigr + bigr + POS + emot. + char n-grams

FS2: $a\Delta(t')n$ + bigr + POS + emot. + char n-grams

Table 3: Results on the Facebook dataset for various TFIDF-weighted features, classification into 2 classes. Macro F-measure, 95% confidence interval $= \pm 0.01$. Underlined numbers show the best results for TFIDF-weighted features. Bold numbers denote the best overall results.

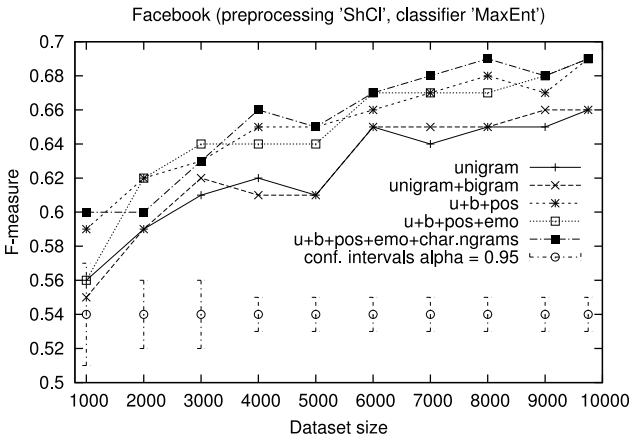

Figure 2: Performance wrt. data size. Using *ShCl* preprocessing and MaxEnt classifier.

fier, all features, *ShCl* preprocessing). Table 4 shows the results as confusion matrices. For each class (*p*—positive, *n*—negative, *0*—neutral) we also report precision, recall, and F-measure. The row headings denote gold labels, the column headings represent values assigned by the annotators or the system.[18] The annotators' results show what can be expected from a 'perfect' system that would solve the task the way a human would.

In general, both annotators judge all three classes with very similar F-measure. By contrast, the system's F-measure is very low for negative posts (0.54 vs. ≈ 0.75 for neutral and positive). We offer the following explanation. First, many of the negative posts surprisingly contain happy emoticons, which

[18]Even though the task has three classes, the annotators also used 'b' for 'bipolar and '?' for 'cannot decide'.

	Annotator 1							
	0	**n**	**p**	**?**	**b**	**P**	**R**	**Fm**
0	4867	136	115	2	54	.93	.94	.93
n	199	1753	6	0	33	.93	.88	.90
p	175	6	2376	0	30	.95	.92	.93
						Macro Fm:		.92
	Annotator 2							
	0	**n**	**p**	**?**	**b**	**P**	**R**	**Fm**
0	4095	495	573	3	8	.95	.79	.86
n	105	1878	6	0	2	.79	.94	.86
p	100	12	2468	3	4	.81	.95	.88
						Macro Fm:		.86
	Best system							
	0	**n**	**p**			**P**	**R**	**Fm**
0	4014	670	490			.74	.78	.76
n	866	1027	98			.57	.52	.54
p	563	102	1922			.77	.74	.75
						Macro Fm:		.69

Table 4: Confusion matrices for three-class classification. 'Best system' configuration: all features (unigram, bigram, POS, emoticons, character n-grams), *ShCl* preprocessing, and MaxEnt classifier. 95% confidence interval $= \pm 0.01$.

could be a misleading feature for the classifier. Second, the language of the negative posts in not as explicit as for the positive ones in many cases; the negativity is 'hidden' in irony, or in a larger context (i.e., *"Now I'm sooo satisfied with your competitor :))"*). This remains an open issue for the future research.

5.2 Product and movie reviews

For the other two datasets, the product reviews and movie reviews, we slightly changed the configuration. First, we removed the character n-grams from the feature sets, otherwise the feature space would become too large for feasible computing. Second, we abandoned SVM as it became computationally infeasible for such a large datasets.

Table 5 (left-hand part) presents results on the product reviews. The combination of unigrams and bigrams works best, almost regardless of the preprocessing. By contrast, POS features rapidly decrease the performance. We suspect that POS features do not carry any useful information in this case and by introducing a lot of 'noise' they cause that the optimization function in the MaxEnt classifier fails to find a global minimum.

In the right-hand part of Table 5 we can see the results on the movie reviews. Again, the bigram feature performs best, paired with combination of HPS stemmer and phonetic transcription (*ShPe*). Adding POS-related features causes a large drop in performance. We can conclude that for larger texts, the bigram-based feature outperforms unigram features and, in some cases, a proper preprocessing may further significantly improve the results.

6 Conclusion

This article presented an in-depth research of supervised machine learning methods for sentiment analysis of Czech social media. We created a large Facebook dataset containing 10,000 posts, accompanied by human annotation with substantial agreement (Cohen's κ 0.66). The dataset is freely available for non-commercial purposes.[19] We thoroughly evaluated various state-of-the-art features and classifiers as well as different language-specific preprocessing techniques. We significantly outperformed the baseline (unigram feature without preprocessing) in three-class classification and achieved F-measure 0.69 using a combination of features (unigrams, bigrams, POS features, emoticons, character n-grams) and preprocessing techniques (unsupervised stemming and phonetic transcription). In addition, we reported results in two other domains (movie and product reviews) with a significant improvement over the baseline.

To the best of our knowledge, this article is the only of its kind that deals with sentiment analysis in Czech social media in such a thorough manner. Not only it uses a dataset that is magnitudes larger than any from the related work, but also incorporates state-of-the-art features and classifiers. We believe that the outcomes of this article will not only help to set the common ground for sentiment analysis for the Czech language but also help to extend the research outside the mainstream languages in this research field.

Acknowledgement

This work was supported by grant no. SGS-2013-029 Advanced computing and information

[19] We encourage other researchers to download our dataset for their research in the sentiment analysis field.

	Product reviews, 3 classes				Movie reviews, 3 classes			
	FS1	FS2	FS3	FS4	FS1	FS2	FS3	FS4
SnCk	0.70	0.74	0.52	0.49	0.76	0.77	0.71	0.61
SnCl	0.71	**0.75**	0.51	0.52	0.76	0.77	0.71	0.70
SlCk	0.67	**0.75**	0.59	0.55	0.78	0.78	0.73	0.72
SlCl	0.67	**0.75**	0.56	0.57	0.78	0.78	0.71	0.71
ShCk	0.67	**0.75**	0.57	0.57	0.78	0.78	0.74	0.72
ShCl	0.67	0.74	0.55	0.57	0.77	0.78	0.73	0.73
SnPe	0.69	0.74	0.50	0.55	0.77	0.78	0.69	0.72
SlPe	0.67	**0.75**	0.55	0.57	0.78	0.78	0.73	0.73
ShPe	0.68	0.74	0.56	0.59	0.78	**0.79**	0.74	0.73
Lp	0.66	**0.75**	0.56	0.57	0.77	0.77	0.68	0.70

Table 5: Results on the product and movie review datasets, classification into 3 classes. FSx denote different feature sets. FS1 = Unigrams; FS2 = Uni + bigrams; FS3 = Uni + big + POS features; FS4 = Uni + big + POS + emot. Macro F-measure, 95% confidence interval ± 0.002 (products), ± 0.003 (movies). Bold numbers denote the best results.

systems and by the European Regional Development Fund (ERDF), project "NTIS - New Technologies for Information Society", European Center of Excellence, CZ.1.05/1.1.00/02.0090. The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005) is highly acknowledged.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

Julie Kane Ahkter and Steven Soria. 2010. Sentiment analysis: Facebook status messages. Technical report, Stanford University. Final Project CS224N.

Alexandra Balahur and Hristo Tanev. 2012. Detecting entity-related events and sentiments from tweets using multilingual resources. In *Proceedings of the 2012 Conference and Labs of the Evaluation Forum Information Access Evaluation meets Multilinguality, Multimodality, and Visual Analytics*.

Ben Blamey, Tom Crick, and Giles Oatley. 2012. R U : -) or : -(? character- vs. word-gram feature selection for sentiment classification of OSN corpora. In *Proceedings of AI-2012, The Thirty-second SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 207–212. Springer.

A. Celikyilmaz, D. Hakkani-Tür, and Junlan Feng. 2010. Probabilistic model-based sentiment analysis of twitter messages. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pages 79–84. IEEE.

Ljiljana Dolamic and Jacques Savoy. 2009. Indexing and stemming approaches for the czech language. *Information Processing and Management*, 45(6):714–720, November.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. 2006. Prague dependency treebank 2.0. Linguistic Data Consortium, Philadelphia.

Mohammad Sadegh Hajmohammadi, Roliana Ibrahim, and Zulaiha Ali Othman. 2012. Opinion mining and sentiment analysis: A survey. *International Journal of Computers & Technology*, 2(3).

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth*

ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '04, pages 168–177, New York, NY, USA. ACM.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*. The AAAI Press.

Gustavo Laboreiro, Luís Sarmento, Jorge Teixeira, and Eugénio Oliveira. 2010. Tokenizing micro-blogging messages using a text classification approach. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, AND '10, pages 81–88, New York, NY, USA. ACM.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, pages 415–463. Springer.

Roque López, Javier Tejada, and Mike Thelwall. 2012. Spanish sentistrength as a tool for opinion mining peruvian facebook and twitter. In *Artificial Intelligence Driven Solutions to Business and Engineering Problems*, pages 82–85. ITHEA, Sofia, Bulgaria.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Justin Martineau and Tim Finin. 2009. Delta TFIDF: An improved feature space for sentiment analysis. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA*. The AAAI Press.

A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña López. 2012. Random walk weighting over sentiwordnet for sentiment polarity detection on twitter. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 3–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*. European Language Resources Association.

Georgios Paltoglou and Mike Thelwall. 2010. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1386–1395, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2013. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Sebastopol, CA 95472.

Josef Steinberger, Polina Lenkova, Mijail Alexandrov Kabadjov, Ralf Steinberger, and Erik Van der Goot. 2011. Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing*, RANLP'11, pages 770–775.

Josef Steinberger, Mohamed Ebrahim, Maud Ehrmann, Ali Hurriyetoglu, Mijail Alexandrov Kabadjov, Polina Lenkova, Ralf Steinberger, Hristo Tanev, Silvia Vázquez, and Vanni Zavarella. 2012. Creating sentiment dictionaries via triangulation. *Decision Support Systems*, 53:689—694.

E.A. Stepanov and G. Riccardi. 2011. Detecting general opinions from customer surveys. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 115–122.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Mikalai Tsytsarau and Themis Palpanas. 2012. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514, May.

Kateřina Veselovská, Jan Hajič Jr., and Jana Šindlerová. 2012. Creating annotated resources for polarity classification in Czech. In *Proceedings of KONVENS 2012*, pages 296–304. ÖGAI, September. PATHOS 2012 workshop.

Liang-Chih Yu, Jheng-Long Wu, Pei-Chann Chang, and Hsuan-Shou Chu. 2013. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge Based Syst*, 41:89–97.

Kunpeng Zhang, Yu Cheng, Yusheng Xie, Daniel Honbo, Ankit Agrawal, Diana Palsetia, Kathy Lee, Wei keng Liao, and Alok N. Choudhary. 2011. SES: Sentiment elicitation system for social media data. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th Conference on, Vancouver, BC, Canada, December 11, 2011*, pages 129–136. IEEE.

Dan Zhang, Luo Si, and Vernon J. Rego. 2012. Sentiment detection with auxiliary data. *Information Retrieval*, 15(3-4):373–390.