Supervised Sentiment Analysis in Czech Social Media[☆]

Ivan Habernal[a,b,*], Tomáš Ptáček[b], Josef Steinberger[a,b]

[a] *Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic*
[b] *NTIS – New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic*

Abstract

This article describes in-depth research on 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 the case of the Czech language no systematic research has yet been 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. We also experiment with five different feature selection algorithms and investigate the influence of named entity recognition and preprocessing on sentiment classification performance. Moreover, in addition to our newly created social media dataset, we also report results for other 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, potentially leading to the production of high-end commercial solutions.

Keywords: sentiment analysis, Czech language, social media, machine learning, feature selection

1. Introduction

Sentiment analysis has become a mainstream research field since the early 2000s. 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). In Czech, there have been very few attempts, although the importance of sentiment analysis of social media became apparent, for example, 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 3.6 million users.[2] One aspect still eludes many of them: automatic analysis of customer

[☆] This article is an extended version of our conference paper (Habernal et al., 2013)

[*] Corresponding author. Tel: +420 377632456
Email addresses: habernal@kiv.zcu.cz (Ivan Habernal), tigi@kiv.zcu.cz (Tomáš Ptáček), jstein@kiv.zcu.cz (Josef Steinberger)

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

[2] http://www.m-journal.cz/cs/jaky-je-skutecny-pocet-ceskych-uzivatelu-facebooku__s288x9161.html [in Czech]

sentiment of products, services, or even a brand or a company name. In many cases, sentiment is still labeled manually, according to 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. Dealing with these issues may lead to novel 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[4] sentiment analysis performed on three different Czech datasets using supervised machine learning. For 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[5] at `http://liks.fav.zcu.cz/sentiment`.

The rest of this article is organized as follows. Section 2 examines the related work with a focus on 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. Section 5 contains the results and provides a thorough discussion. Finally, Section 5.3 explores the influence of feature selection methods.

2. Related work

There are two basic approaches to sentiment analysis: dictionary-based and machine learning-based. Whereas 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 the 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 improvements in delta TFIDF weighting.

The focus of 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 a 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 (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 the presence of emoticons (usually positive and negative) and are then filtered manually. Although large high-quality collections can be created very quickly with this approach, it makes a strong assumption that every positive or negative post must contain an emoticon.

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

[4] Or *post-level*, as documents correspond to *posts* in social media.

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

Balahur and Tanev (2012) performed experiments with Twitter posts as part of the CLEF 2012 RepLab.[6] They classified English and Spanish tweets with a small but precise lexicon, which also contained slang, combined with a set of rules that captured the manner in which sentiment is expressed in social media.

Finally, we would like to direct the reader to an in-depth survey by Tsytsarau and Palpanas (2012) for actual results obtained from the above-mentioned methods.

2.2. Feature selection for improving classification performance

The basic reason for using feature selection (or reduction) methods for supervised sentiment analysis is twofold: first, the reduced feature set decreases the computing demands for the classifier, and, second, removing irrelevant features can lead to better classification accuracy. Furthermore, noise and redundancy in the feature space increase the likelihood of overfitting (Abbasi et al., 2011).

A study by Sharma and Dey (2012) compares five methods for feature selection, namely Information Gain, Chi Square, Gain Ratio, Relief-F, and Document Frequency, together with seven different classifiers. Results are reported on the widely-used movie review database from Pang et al. (2002). The best performance was achieved by using the SVM classifier and the Gain Ratio selector with the number of features ranging from 2,000 to 8,000 and employing only unigrams as features sorted by their frequency.

Abbasi et al. (2008) proposed an entropy-weighted genetic algorithm that combines Information Gain with a genetic algorithm for selecting features in a bootstrapping manner, tuned on held-out data. They performed document-level binary sentiment of English and Arabic and used SVM as the main classifier. Their results were superior to other approaches, such as plain SVM or Information Gain selection. In their later work, Abbasi et al. (2011) proposed another feature selection method called the Feature Relation Network. This manually constructed network of feature dependencies (e.g., subsumption[7] or parallel relations of various n-grams) relies on SentiWordNet in order to assign the final feature weights.

One of the classical papers on feature selection for text classification by Forman (2003) proposes a metric called Bi-Normal Separation and provides an extensive comparison with another twelve existing feature selection methods. Using SVM as the underlying classifier, the proposed method yields the best results and is suitable for skewed (imbalanced) classes. Other examples of feature selection methods for sentiment analysis or text classification can be found in, e.g., (Chen et al., 2009; Aghdam et al., 2009).

Since feature selection is also important outside the domain of text classification, Wasikowski and Chen (2010) conducted a systematic study, focusing on dealing with class imbalance on small samples. They compare seven selection methods on 11 small datasets with highly skewed classes and conclude by recommending two best-performing algorithms, especially for scenarios that require a small number of features. Another approach based on dynamic mutual information is presented in (Liu et al., 2009). Again, the experiments are conducted on 16 benchmark datasets with a rather small size (up to 8124 instances only) and a small number of features (from 18 to 279), which is a fundamentally different scenario from machine learning-based sentiment analysis.

Feature selection, however, does not have to lead to a better performance in all cases, as reported e.g. by Boiy and Moens (2009), who report Chi-square selection results in their preliminary tests without any success.

2.3. Sentiment analysis in the 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, and because of its small size 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 a third language by means of a

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

[7] *'is-a'* hierarchical relation

state-of-the-art machine translation service. Finally, the resulting sentiment dictionaries were merged using the overlap of the two automatic translations.

A multilingual parallel news corpus annotated with opinions on 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 seven languages including Czech. The research targets fundamentally different objectives from our research as it focuses on news media and aspect-based sentiment analysis.

Recent experiments with incorporating word clusters as additional features to tackle the issue of the high flection of Czech were successful, especially in the Czech movie review domain (Habernal and Brychcín, 2013). The clusters were obtained from semantic spaces created on unlabeled external data. Further improvement of the overall classification performance was achieved by exploiting the global target of the analyzed reviews using Gibbs sampling (Brychcín and Habernal, 2013).

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.[8] 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[9] 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 marketing feedback on 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 constitute negative sentiment, whereas joy or happiness about the brand is treated as positive. We also added another class called *bipolar* which represents both positive and negative sentiment in one post.[10] In some cases, the user's opinion, although positive, does not relate to the given page.[11] Therefore the sentiment is treated as neutral in these cases, in accordance with our above-mentioned assumption.

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

The gold data were 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. Even after the third annotator's labeling, however, 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 of previous posts. These issues remain open.

The Facebook dataset contains 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 source pages is shown in Figure 1. The statistics reveal negative opinions towards cell phone operators (`www.facebook.com/o2cz`, `www.facebook.com/TmobileCz`, and `www.facebook.com/vodafoneCZ`) and positive opinions towards, e.g., perfume sellers (`www.facebook.com/Xparfemy.cz`) and Prague Zoo (`www.facebook.com/zoopraha`).

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

[9] `http://code.google.com/p/jlangdetect/`

[10] 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"*.

[11] Certain campaigns ask the fans to, e.g., write a poem—these posts are mostly positive (or funny, at least) but are irrelevant in terms of the desired task.

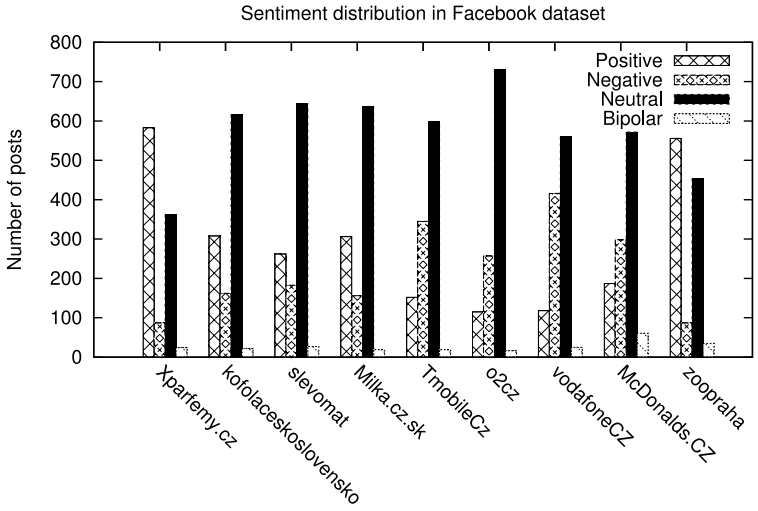


Figure 1: Social media dataset statistics

3.2. Movie review dataset

Movie reviews as a corpus for sentiment analysis have 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[12] and split them into three categories according to their star rating (0–2 stars as negative, 3–4 stars as neutral, 5–6 stars as positive). The dataset contains 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 scraped all user reviews from a large Czech e-shop Mall.cz[13] which offers a wide range of products. The product reviews are accompanied by star ratings on a scale of zero to five. 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 four-star ratings mostly correspond to neutral opinions and three or fewer 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 the 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 of social media. Furthermore, we remove stopwords using the stopword list from Apache Lucene project.[14]

[12] `http://www.csfd.cz/`
[13] `http://www.mall.cz`
[14] `http://lucene.apache.org/core/`

In many NLP applications, a very popular preprocessing technique is stemming. We tested the Czech light stemmer (Dolamic and Savoy, 2009) and High Precision Stemmer.[15] Another widely-used method for reducing the vocabulary size, and thus the feature space, is lemmatization. For the Czech language the only currently available lemmatizer is shipped with the Prague Dependency Treebank (PDT) toolkit (Hajič et al., 2006). We, however, used our in-house Java HMM-based implementation with the PDT training data as we needed better control over each preprocessing step. Kanis and Skorkovská (2010) experimented with a lemmatizer based on *Ispell*. Following their work, we developed an in-house lemmatizer using rules and dictionaries from the *OpenOffice* suite.

Part-of-speech tagging was done with our in-house Java solution that utilizes Prague Dependency Treebank (PDT) data as well. Since, however, PDT is trained on news corpora, we doubt it is suitable for tagging social media that are written in very informal language (see, e.g., (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 the International Phonetic Alphabet (IPA) in order to reduce the effect of these mistakes. We relied on eSpeak[16] implementation. Another preprocessing step might involve removing diacritics, as many Czech users type only unaccented characters. Posts without diacritics, however, represent only about 8% of our datasets, and thus we decided to keep diacritics intact.

We were also interested in whether named entities (e.g., product names, brands, places, etc.) carry sentiment and how their presence influences classification accuracy. For these experiments, we employ a CRF-based named entity recognizer (Konkol and Konopík, 2013) and replace the words identified as entities with their respective entity type (e.g., *McDonald's* becomes *company*). This preprocessing has not been widely discussed in the literature devoted to document-level sentiment analysis, but Boiy and Moens (2009), for example, remove the 'entity of interest' in their approach.

The complete preprocessing diagram and its variants is depicted in Table 1. Overall, there are 16 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 empirically set to five. 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 five, 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 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 the number of negative verbs obtained from POS tags.

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

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

[16] `http://espeak.sourceforge.net`

Pipe 1	Pipe 2	Pipe 3
Tokenizing		
ArkTweetNLP		
POS tagging		
PDT		
Named entity filtering (N) [optional]		
remove (r)		
Stem (S)		**Lemma (L)**
none (n)		PDT (p)
light (l)		OpenOffice (o)
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 "NrLp" means *1. tokenizing, 2. POS-tagging, 3. removing named entities, 4. lemmatization using PDT, and 5. removing stopwords.*

Delta TFIDF variants for binary scenarios. Although simple binary word features (presence of a certain word) achieve a surprisingly good performance, they have been surpassed by various TFIDF-based weightings, such as Delta TFIDF (Martineau and Finin, 2009), and Delta BM25 TFIDF (Paltoglou and Thelwall, 2010). Delta-TFIDF still uses traditional TFIDF word weighting but treats positive and negative documents differently. All the existing related works which use this kind of feature, however, deal only with binary decisions (positive/negative), and thus we filtered out neutral documents from the datasets.[17] 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. Feature Selection

Feature selection methods assign a certain weight to each feature, depending on its significance (discriminative power) for each class. After the weights are obtained, the top k features can be kept for the classifier, or the features with low weight can be cut off at a certain threshold.

Let t_k and $\bar{t_k}$ denote the presence and absence, respectively, of a particular feature in a certain class (e.g., c_1, and c_2). To estimate the joint probability of a feature in a given class, we will use the following table with appropriate feature counts:

	c_1	c_2
t_k	a	b
$\bar{t_k}$	c	d

Then N denotes the total number of features in all classes, $N = a + b + c + d$. The joint probability $p(t_k, c_1)$ can then be estimated as

$$p(t_k, c_1) = \frac{a}{N}, \tag{1}$$

and similarly for $p(t_k, c_2)$. The probability of a particular feature in all classes $p(t_k)$ is given by

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

$$p(t_k) = \frac{a+b}{N}. \tag{2}$$

Furthermore, c_1 can be estimated as

$$p(c_1) = \frac{a+c}{N}. \tag{3}$$

The conditional probability of t_k given c_1 is given by

$$p(t_k|c_1) = \frac{a}{a+c}. \tag{4}$$

Henceforth, let n denote the number of classes, $m = \{t_k, \bar{t_k}\}$, and all logarithms are to the base 2.

We follow with the formulas for the particular feature selection methods. For a more detailed discussion of these methods, please refer to, e.g., (Forman, 2003; Zheng et al., 2004; Uchyigit, 2012; Patočka, 2013).

4.3.1. Mutual Information (MI)

Mutual Information is always non-negative and symmetrical, $MI(X,Y) = MI(Y,X)$.

$$MI = \sum_{i=0}^{n}\sum_{k=0}^{m} \log \frac{p(t_k, c_i)}{p(c_i)p(t_k)} \tag{5}$$

4.3.2. Information Gain (IG)

Also known as *Kullback-Leibler divergence* or *relative entropy*. It is a non-negative and asymmetrical metric.

$$IG = \sum_{i=0}^{n}\sum_{k=0}^{m} p(t_k, c_i) \log \frac{p(t_k, c_i)}{p(c_i)p(t_k)} + p(\bar{t_k}, c_i) \log \frac{p(\bar{t_k}, c_i)}{p(c_i)p(\bar{t_k})} \tag{6}$$

4.3.3. Chi Square (CHI)

Chi Square (χ^2) can be derived as follows.

$$GSS(t_k, c_i) = p(t_k, c_i)p(\bar{t_k}, \bar{c_i}) - p(t_k, \bar{c_i})p(t_k, \bar{c_i}), \tag{7}$$

$$NGL(t_k, c_i) = \frac{\sqrt{N} \cdot GSS(t_k, c_i)}{\sqrt{p(t_k)p(\bar{t_k})p(c_i)p(\bar{c_i})}}, \tag{8}$$

$$\chi^2 = \sum_{i=0}^{n}\sum_{k=0}^{m} NGL(t_k, c_i)^2 \tag{9}$$

4.3.4. Odds Ratio (OR)

$$OR = \sum_{i=0}^{n}\sum_{k=0}^{m} \log \frac{p(t_k|c_i)p(\bar{t_k}|\bar{c_i})}{p(\bar{t_k}|c_i)p(t_k|\bar{c_i})} \tag{10}$$

4.3.5. Relevancy Score (RS)

$$RS = \sum_{i=0}^{n}\sum_{k=0}^{m} \log \frac{p(t_k|c_i)}{p(\bar{t_k}|\bar{c_i})} \tag{11}$$

4.4. Classifiers

All evaluation tests were performed with two classifiers, Maximum Entropy (MaxEnt) and Support Vector Machines (SVM). Although the Naive Bayes classifier is also widely used in 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[18] with default settings (the 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 scenario, we classified into all three classes (positive, negative, and neutral).[19] In the second scenario, we followed 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 the MaxEnt classifier into the second scenario.

All tests were conducted in the 10-fold cross validation manner. We report the macro F-measure, as it allows comparison of classifier results on different datasets. Moreover, we do not report the 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 three-class classification scenario on the Facebook dataset. The row labels denote the preprocessing configuration according to Table 1. In most cases, the maximum entropy classifier significantly outperforms SVM. The combination of all features (the last column) yields the best results regardless of the preprocessing steps. The reason might be that the 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 FS4: *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. Because of the 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 the TFIDF-weighted feature $a\Delta(t')n$. It turned out that the weighted form of the word feature does not improve the performance, compared with the simple binary unigram feature. Furthermore, a set of different features (words, bigrams, POS, emoticons, character n-grams) significantly outperforms a single TFIDF-weighted feature.

Furthermore, we 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

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

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

Facebook dataset, 3 classes

Feat. set	FS1		FS2		FS3		FS4		FS5	
Classifier	ME	SVM	ME	SVM	ME	SVM	ME	SVM	ME	SVM
SnCk	63	64	63	64	66	64	66	64	**69**	67
SnCl	63	64	63	64	66	63	66	63	**69**	68
SlCk	65	67	66	67	68	66	67	66	**69**	67
SlCl	65	67	65	67	68	66	**69**	66	**69**	67
ShCk	66	67	66	67	68	67	67	67	**69**	67
ShCl	66	66	66	67	**69**	67	**69**	67	**69**	67
SnPe	64	65	64	65	67	65	67	65	68	68
SlPe	65	67	65	67	68	67	67	66	68	67
ShPe	66	67	66	67	**69**	66	**69**	66	68	67
Lp	64	65	63	65	67	64	67	65	68	67
Lo	65	66	64	66	67	66	67	65	68	67

Table 2: Results for the Facebook dataset, classification into three classes. Macro F-measure (in %), 95% confidence interval $\approx \pm 1$. Bold numbers denote the best results. **FS1**: Unigrams; **FS2**: unigrams, bigrams; **FS3**: unigrams, bigrams, POS features; **FS4**: unigrams, bigrams, POS, emoticons; **FS5**: unigrams, bigrams, POS, emoticons, character n-grams.

2, while the dataset grows to approx 6k to 7k items, the performance rises for most combinations of features. With a 7k-item 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.

The influence of named entity filtering is shown in Table 4. In most cases, removing named entities leads to a significant drop in classification. Thus we can conclude that in our corpus, the named entities themselves represent an important opinion-holder. This also corresponds to the sentiment distribution as shown in Figure 1 (e.g., sentiment towards mobile phone operators is rather negative) and thus by removing the brand name from the data the classifier loses useful information.

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 had to be solved by a third or even a fourth person. Thus even the original annotators do not achieve a 1.00 F-measure on the gold data.

We present 'performance' results of both annotators and of the best system as well (MaxEnt classifier, all features, *ShCl* preprocessing). Table 5 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.[20] The annotators' results show what can be expected from a 'perfect' system that solves the task the way a human would.

In general, both annotators judge all three classes with very similar F-measures. 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 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 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

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

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	83	86	86	86	85	86	86	**90**	89
SnCl	84	86	86	86	86	86	86	**90**	**90**
various*	85	<u>88</u>	<u>88</u>	<u>88</u>	<u>88</u>	<u>88</u>	<u>88</u>	**90**	**90**
SnPe	84	86	86	86	86	86	86	**90**	**90**
Lp	84	86	85	85	86	86	86	88	88
Lo	84	88	87	87	87	87	87	**90**	**90**

* 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 for the Facebook dataset for various TFIDF-weighted features, classification into two classes. Macro F-measure (in %), 95% confidence interval $\approx \pm 1$. Underlined numbers show the best results for TFIDF-weighted features. Bold numbers denote the best overall results.

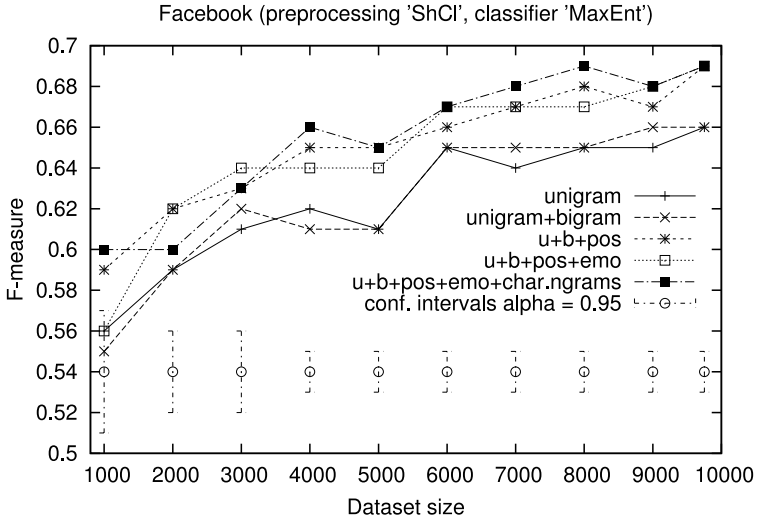


Figure 2: Learning curve; using *ShCl* preprocessing and MaxEnt classifier.

Facebook dataset, 3 classes

Feat. set	FS1		FS2		FS3		FS4		FS5	
Classifier	ME	SVM	ME	SVM	ME	SVM	ME	SVM	ME	SVM
SlPe	65	67	65	67	68	67	67	66	68	67
NrSlPe	65	66	65	65	67	65	68	65	68	66
SlCl	65	67	65	67	68	66	**69**	66	**69**	67
NrSlCl	65	66	65	66	67	65	68	65	68	67
SlCk	65	67	66	67	68	66	67	66	**69**	67
NrSlCk	65	66	65	66	67	65	67	65	68	67
ShPe	66	67	66	67	**69**	66	**69**	66	68	67
NrShPe	65	66	65	66	67	65	68	65	67	66
ShCl	66	66	66	67	**69**	67	**69**	67	**69**	67
NrShCl	65	66	65	66	67	66	68	64	68	66

Table 4: Comparison of the five best (on average) preprocessing pipes with and without NER (Nr prefix). Results for the Facebook dataset, classification into three classes. Macro F-measure (in %), 95% confidence interval $\approx \pm 1$. Bold numbers denote the best results. **FS1**: Unigrams; **FS2**: unigrams, bigrams; **FS3**: unigrams, bigrams, POS features; **FS4**: unigrams, bigrams, POS, emoticons; **FS5**: unigrams, bigrams, POS, emoticons, character n-grams.

too large for feasible computing. Second, we abandoned SVM as it became computationally infeasible for such large datasets.

Table 6 (left-hand side) presents the results of 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-related features do not carry any useful information in this case and also bring too much 'noise' to the classifier.

In the right-hand side of Table 6 we can see the results of the movie reviews. Again, the bigram feature performs best, paired with a 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.

Table 7 shows the effect of replacing named entities by their types. Again, named entities (e.g., actors, directors, products, brands) are very strong opinion-holders and thus their filtering significantly decreases classification performance.

5.3. Feature selection experiments

Using the two most promising preprocessing pipelines (*ShCl*, *ShPe*), we conducted experiments with feature selection methods as introduced in Section 4.3. We classify into three classes using both MaxEnt and SVM classifiers on the Facebook dataset and using only MaxEnt on the other datasets (because of computational feasibility, as mentioned previously in Section 5.2).

Feature selection methods assign a certain weight to each feature and cut off those features whose weight is under a certain threshold. To estimate an optimal parameter automatically, we measured how the feature weight threshold influences the performance. For this purposes we used held-out data (10% of the training data). In each fold of the 10-fold cross validation, the optimum threshold for feature cut-off was set such that the performance on the held-out data was maximized.

In the previous experiments (Section 5), the feature space was pruned by a minimum occurrence which was empirically set to five. This prior pruning is not necessary for automatic feature selection. Therefore, we removed this prior filtering for the experiments on the Facebook data.[21]

[21]For the other two datasets, the product reviews and movie reviews, we still kept the minimum occurrence set to five, as otherwise the feature space would become too large for feasible computing.

	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

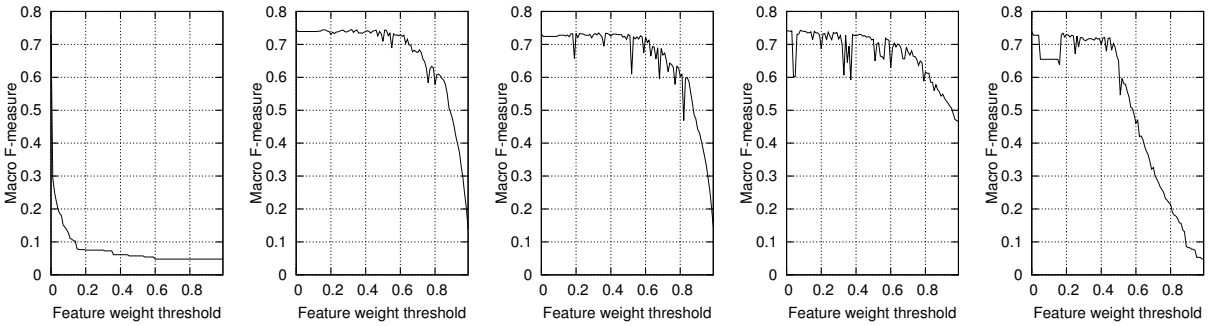
Table 5: Confusion matrices for three-class classification. 'Best system' configuration: all features (unigram, bigram, POS, emoticons, character n-grams), *ShCl* preprocessing, and MaxEnt classifier. 95% confidence interval $\approx \pm 1$.

	Product reviews, 3 classes				Movie reviews, 3 classes			
	FS1	FS2	FS3	FS4	FS1	FS2	FS3	FS4
SnCk	69.90	74.00	52.41	49.02	75.94	77.02	70.72	61.44
SnCl	70.79	75.05	50.93	51.73	76.03	77.15	70.60	69.70
SlCk	66.87	**75.18**	58.52	55.49	77.92	**78.26**	73.25	72.09
SlCl	67.26	74.74	56.48	56.99	77.60	**78.35**	70.77	71.23
ShCk	66.90	74.68	57.39	56.91	77.82	78.23	73.80	71.59
ShCl	66.83	74.02	54.88	57.43	77.06	78.21	73.14	73.16
SnPe	69.42	74.20	50.01	55.46	76.59	77.67	69.27	72.50
SlPe	66.70	**75.23**	55.08	57.03	77.60	**78.26**	72.94	73.22
ShPe	67.54	73.38	56.22	59.47	77.62	**78.50**	73.86	72.68
Lp	65.60	74.68	56.18	56.68	76.94	77.01	67.87	69.80
Lo	68.11	**75.30**	52.83	54.03	76.17	77.37	72.93	72.04

Table 6: Results for the product and movie review datasets, classification into three classes. FSx denote different feature sets. **FS1** = Unigrams; **FS2** = Uni + bigrams; **FS3** = Uni + big + POS features; **FS4** = Uni + big + POS + emot. Macro F-measure (in %), 95% confidence interval $\approx \pm 0.2$ (products), $\approx \pm 0.3$ (movies). Bold numbers denote the best results.

	Product reviews, 3 classes				Movie reviews, 3 classes			
	FS1	FS2	FS3	FS4	FS1	FS2	FS3	FS4
SlCk	66.87	**75.18**	58.52	55.49	77.92	**78.26**	73.25	72.09
NrSlCk	66.54	72.57	50.39	56.66	75.91	75.98	67.84	70.47
SlCl	67.26	74.74	56.48	56.99	77.60	**78.35**	70.77	71.23
NrSlCl	66.54	72.57	50.39	52.36	75.91	75.98	67.84	70.99
ShCl	66.83	74.02	54.88	57.43	77.06	78.21	73.14	73.16
NrShCl	66.19	71.94	56.13	58.91	75.79	75.86	72.99	72.39
SlPe	66.70	**75.23**	55.08	57.03	77.60	**78.26**	72.94	73.22
NrSlPe	64.98	74.45	49.39	55.97	76.09	76.07	72.09	68.33
ShPe	67.54	73.38	56.22	59.47	77.62	**78.50**	73.86	72.68
NrShPe	66.60	74.33	55.00	56.10	76.15	76.26	70.88	71.62

Table 7: Comparison of the five best (on average) preprocessing pipes with and without NER (Nr prefix). Results for the product and movie review datasets, classification into three classes. FSx denotes different feature sets. **FS1** = Unigrams; **FS2** = Uni + bigrams; **FS3** = Uni + big + POS features; **FS4** = Uni + big + POS + emot. Macro F-measure (in %), 95% confidence interval $\approx \pm 0.2$ (products), $\approx \pm 0.3$ (movies). Bold numbers denote the best results.



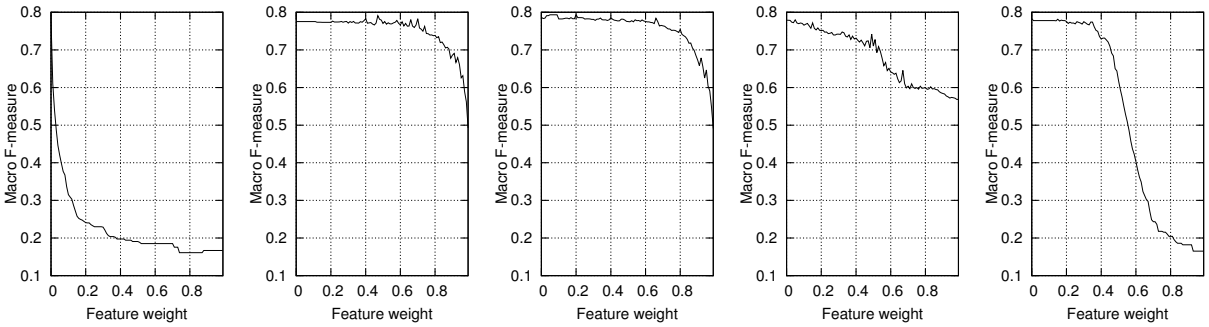
(a) Chi Square (b) Information Gain (c) Mutual Inf. (d) Odds Ratio (e) Relevancy Score

Figure 3: Feature weight threshold estimation on heldout data. Product reviews, *ShCl* preprocessing pipe, *MaxEnt* classifier, 3 classes, **FS2:** unigrams, bigrams

Figures 3, 4, 5, and 6 show dependency graphs of the macro F-measure given a feature weight threshold. Note that these figures depict parameter estimation for only one fold from the 10-fold cross-validation and thus serve only as an illustration of the feature selection behavior. It is apparent that *Information Gain* and *Mutual Information* are able to filter out noisy features to a large extent yet keep the performance almost unchanged. The worst selector is *Chi Square* as it drastically lowers the performance even with a very small filtering threshold.

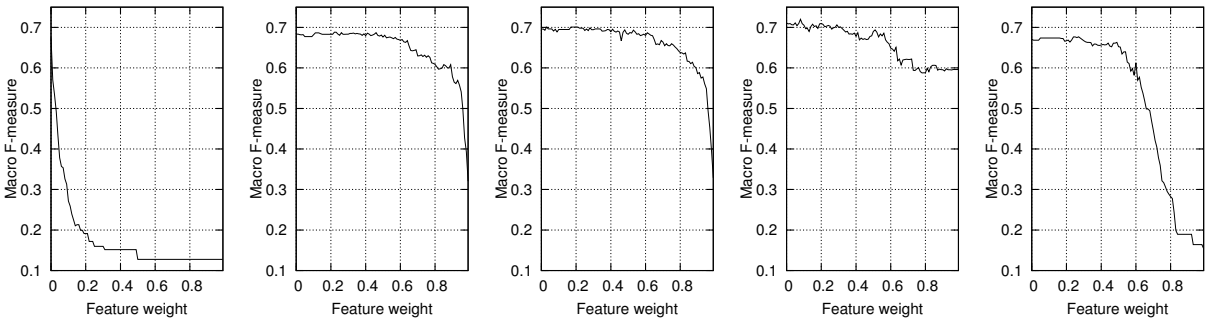
Overall, a significant improvement from 73.38% (baseline) to 73.85% was achieved for the product reviews, by means of the *Mutual Information* feature selector and *ShPe* preprocessing pipeline (see Table 8). Yet very similar results were obtained with a different preprocessing pipeline (*ShCl*). For the movie reviews dataset (Table 9) and the Facebook dataset with and without feature space pruning (Tables 10, and 11, respectively) no significant improvement was achieved.

We can conclude that, in our settings, feature selection does not lead to a better overall performance, however, it can speed up the classification by filtering out noisy features.



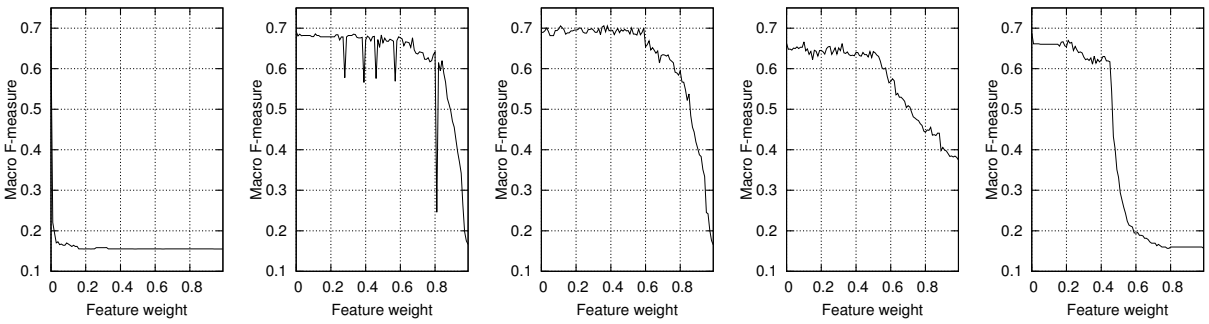
(a) Chi Square (b) Information Gain (c) Mutual Inf. (d) Odds Ratio (e) Relevancy Score

Figure 4: Feature weight threshold estimation using heldout data. Movie reviews, *ShCl* preprocessing pipe, *MaxEnt* classifier, three classes, **FS2:** unigrams, bigrams



(a) Chi Square (b) Information Gain (c) Mutual Inf. (d) Odds Ratio (e) Relevancy Score

Figure 5: Feature weight threshold estimation using heldout data. Facebook dataset, *ShCl* preprocessing pipe, *MaxEnt* classifier, three classes, no prior feature space pruning, **FS5:** unigrams, bigrams, POS, emoticons, character n-grams



(a) Chi Square (b) Information Gain (c) Mutual Inf. (d) Odds Ratio (e) Relevancy Score

Figure 6: Feature weight threshold estimation using heldout data. Facebook dataset, *ShCl* preprocessing pipe, *SVM* classifier, three classes, no prior feature space pruning, **FS5:** unigrams, bigrams, POS, emoticons, character n-grams

Product reviews, 3 classes

Feat. selection	ChS		IG		MI		OR		RS		-	
Feat. set	FS1	FS2	FS1	FS2	FS1	FS2	FS1	FS2	FS1	FS2	FS1	FS2
ShCl	65.93	73.50	66.28	**73.61**	66.16	73.56	65.50	72.74	66.03	73.40	66.83	**74.02**
ShPe	65.93	72.54	66.19	73.02	66.63	**73.85**	66.53	71.94	66.01	72.76	67.54	73.38

Table 8: Results for the product review dataset, classification into three classes. Macro F-measure (in %), 95% confidence interval $\approx \pm 0.2$. Bold numbers denote the best results. **FS1**: Unigrams; **FS2**: unigrams, bigrams.

Movie reviews, 3 classes

Feat. selection	ChS		IG		MI		OR		RS		-	
Feat. set	FS1	FS2	FS1	FS2	FS1	FS2	FS1	FS2	FS1	FS2	FS1	FS2
ShCl	76.86	**78.46**	77.37	78.03	76.98	77.44	76.18	77.71	77.32	77.43	77.06	**78.21**
ShPe	77.43	77.83	76.64	77.85	76.77	77.62	75.89	78.06	77.25	77.38	77.62	**78.50**

Table 9: Results for the movie review dataset, classification into three classes. Macro F-measure (in %), 95% confidence interval $\approx \pm 0.3$. Bold numbers denote the best results. **FS1**: Unigrams; **FS2**: unigrams, bigrams.

Facebook dataset, 3 classes

	Feat. set	FS1		FS2		FS3		FS4		FS5	
Feat. selection	Classifier	ME	SVM	ME	SVM	ME	SVM	ME	SVM	ME	SVM
ChS	ShCl	64.88	66.36	65.43	67.09	67.32	65.38	**68.51**	65.70	**68.63**	67.16
	ShPe	65.68	67.13	65.04	66.95	68.36	65.90	67.57	66.18	67.46	66.25
IG	ShCl	64.39	65.50	65.54	66.24	67.64	65.64	67.42	65.90	**68.56**	66.23
	ShPe	64.18	66.04	65.18	66.20	67.72	65.36	67.53	64.74	67.96	66.30
MI	ShCl	64.40	66.08	64.37	65.45	67.94	64.38	67.43	65.77	**68.73**	66.90
	ShPe	64.05	66.39	64.30	66.10	67.63	65.82	68.22	65.42	67.50	65.91
OR	ShCl	64.68	66.10	65.31	66.91	67.77	65.66	67.03	64.16	67.24	66.84
	ShPe	64.77	66.79	64.31	66.51	67.94	64.12	67.55	65.60	66.70	66.03
RS	ShCl	64.68	65.80	65.28	66.32	67.72	65.05	67.44	65.67	68.13	66.14
	ShPe	63.90	65.96	64.83	66.75	66.98	65.66	67.05	64.71	67.99	66.49
-	ShCl	65.69	66.26	65.73	66.89	**68.85**	66.75	**68.96**	67.06	**68.76**	66.71
	ShPe	65.74	66.76	65.66	66.95	**68.72**	65.81	**68.66**	66.05	68.45	66.57

Table 10: Results for the Facebook dataset, classification into three classes. Macro F-measure (in %), 95% confidence interval $\approx \pm 1$. Bold numbers denote the best results. **FS1**: Unigrams; **FS2**: unigrams, bigrams; **FS3**: unigrams, bigrams, POS features; **FS4**: unigrams, bigrams, POS, emoticons; **FS5**: unigrams, bigrams, POS, emoticons, character n-grams.

Facebook dataset, 3 classes, no prior feature space pruning

	Feat. set	FS1		FS2		FS3		FS4		FS5	
Feat. selection	Classifier	ME	SVM	ME	SVM	ME	SVM	ME	SVM	ME	SVM
ChS	ShCl	65.57	67.81	67.28	66.64	67.60	65.12	67.76	64.37	**69.04**	66.97
	ShPe	65.91	68.04	67.19	67.43	67.33	64.96	67.66	64.99	68.31	66.21
IG	ShCl	65.55	67.55	67.69	66.48	66.64	64.28	66.91	64.52	**69.01**	66.87
	ShPe	64.84	67.62	67.27	66.59	67.57	65.08	67.10	64.78	68.17	66.19
MI	ShCl	65.10	66.99	67.51	66.84	67.05	64.94	68.07	64.22	**69.35**	66.84
	ShPe	65.44	66.96	67.40	66.84	67.11	64.34	67.74	64.50	68.14	66.69
OR	ShCl	66.09	67.25	67.61	66.67	67.68	64.44	66.71	64.84	**69.27**	67.20
	ShPe	65.03	68.07	67.37	66.23	67.19	65.26	67.38	65.62	68.14	65.82
RS	ShCl	64.78	67.89	67.34	67.18	66.83	67.11	67.51	64.21	68.48	66.21
	ShPe	65.37	67.55	67.37	66.26	67.61	67.10	67.42	64.33	67.59	66.14
-	ShCl	65.21	68.55	67.85	66.97	68.10	65.48	68.33	65.01	**69.37**	67.46
	ShPe	66.09	68.27	67.65	67.44	68.11	65.94	67.27	65.90	68.57	66.90

Table 11: Results for the Facebook dataset, classification into three classes. Macro F-measure (in %), 95% confidence interval $\approx \pm 1$. Bold numbers denote the best results. **FS1**: Unigrams; **FS2**: unigrams, bigrams; **FS3**: unigrams, bigrams, POS features; **FS4**: unigrams, bigrams, POS, emoticons; **FS5**: unigrams, bigrams, POS, emoticons, character n-grams.

5.4. Summary of results for social media

Given the results achieved on the Facebook dataset, the following strategies for sentiment analysis of social media in Czech can be considered. First, the preprocessing pipeline should take into account text properties typical of social media, such as proper tokenization (with respect to emoticons, URLs, etc.), stemming, and lower-casing. Additional normalization, such as phonetic transcription, can also increase performance because of the many grammatical errors present in such texts (the case of, e.g., *i/y; ie/ě* in Czech). Second, the Maximum Entropy classifier yields better results than the linear kernel SVM; moreover, training is significantly shorter. The feature set consisting of unigrams, bigrams, emoticons, and various POS features gives the best overall results. Third, filtering named entities or feature selection did not improve the overall performance for our dataset.

It should be noted that the number of examined domains was limited because we restricted our dataset only to the top nine most popular Czech Facebook brand pages. It is thus worth investigating how the system would tackle the issues of domain-dependent features and domain portability. This remains an open question for future work.

6. Conclusion

This article presented in-depth research on 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.[22] We thoroughly evaluated various state-of-the-art features and classifiers as well as different language-specific preprocessing techniques and feature selection algorithms. We significantly outperformed the baseline (unigram feature without preprocessing) in three-class classification and achieved an F-measure of 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 first one of its kind that deals with sentiment analysis in Czech social media in such a thorough manner. Not only does it use a dataset that is magnitudes larger

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

than any in 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 beyond the mainstream languages and may be applied to sentiment analysis in other Slavic languages, such as Slovak or Polish.

Acknowledgments

This work was supported by grant no. SGS-2013-029 Advanced computing and information systems, by a POSTDOC grant from the University of West Bohemia, 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. 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 gratefully acknowledged. Access to the CERIT-SC computing and storage facilities provided under the programme Center CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, reg. no. CZ. 1.05/3.2.00/08.0144, is greatly appreciated. We thank Tomáš Brychcín for his High-Precision Stemmer implementation, Michal Konkol for his machine learning library, and Michal Patočka for his implementation of feature selection algorithms.

References

Abbasi, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Trans. Inf. Syst. 26, 12:1–12:34. doi:10.1145/1361684.1361685.

Abbasi, A., France, S., Zhang, Z., Chen, H., 2011. Selecting attributes for sentiment classification using feature relation networks. IEEE Transactions on Knowledge and Data Engineering 23, 447–462.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R., 2011. Sentiment analysis of twitter data, in: Proceedings of the Workshop on Languages in Social Media, Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 30–38.

Aghdam, M.H., Ghasem-Aghaee, N., Basiri, M.E., 2009. Text feature selection using ant colony optimization. Expert Syst. Appl. 36, 6843–6853.

Ahkter, J.K., Soria, S., 2010. Sentiment analysis: Facebook status messages. Technical Report. Stanford University. Final Project CS224N.

Balahur, A., Tanev, H., 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.

Blamey, B., Crick, T., Oatley, G., 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, Springer. pp. 207–212.

Boiy, E., Moens, M.F., 2009. A machine learning approach to sentiment analysis in multilingual web texts. Information Retrieval 12, 526–558. doi:10.1007/s10791-008-9070-z.

Brychcín, T., Habernal, I., 2013. Unsupervised improving of sentiment analysis using global target context, in: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria. pp. 122–128. URL: http://www.aclweb.org/anthology/R13-1016.

Celikyilmaz, A., Hakkani-Tür, D., Feng, J., 2010. Probabilistic model-based sentiment analysis of twitter messages, in: Spoken Language Technology Workshop (SLT), 2010 IEEE, IEEE. pp. 79–84.

Chen, J., Huang, H., Tian, S., Qu, Y., 2009. Feature selection for text classification with naïve bayes. Expert Syst. Appl. 36, 5432–5435.

Dolamic, L., Savoy, J., 2009. Indexing and stemming approaches for the czech language. Information Processing and Management 45, 714–720.

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3, 1289–1305.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A., 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, Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 42–47.

Go, A., Bhayani, R., Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford.

Habernal, I., Brychcín, T., 2013. Semantic spaces for sentiment analysis, in: Habernal, I., Matoušek, V. (Eds.), Proceedings of the 16th international conference on Text, Speech and Dialogue (TSD'13). Springer, Berlin/Heidelberg. volume 8082 of *Lecture Notes in Computer Science*, pp. 484–491.

Habernal, I., Ptáček, T., Steinberger, J., 2013. Sentiment analysis in czech social media using supervised machine learning, in: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Atlanta, Georgia. pp. 65–74. URL: http://www.aclweb.org/anthology/W13-1609.

Hajič, J., Panevová, J., Hajičová, E., Panevová, J., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., 2006. Prague dependency treebank 2.0. Linguistic Data Consortium, Philadelphia.

Hajmohammadi, M.S., Ibrahim, R., Othman, Z.A., 2012. Opinion mining and sentiment analysis: A survey. International Journal of Computers & Technology 2.

Hu, M., Liu, B., 2004. Mining and summarizing customer reviews, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA. pp. 168–177.

Kanis, J., Skorkovská, L., 2010. Comparison of different lemmatization approaches through the means of information retrieval performance, in: Sojka, P., Horák, A., Kopeček, I., Pala, K. (Eds.), Text, Speech and Dialogue. Springer Berlin Heidelberg. volume 6231 of *Lecture Notes in Computer Science*, pp. 93–100.

Konkol, M., Konopík, M., 2013. CRF-based czech named entity recognizer and consolidation of czech NER research, in: Habernal, I., Matoušek, V. (Eds.), Proceedings of the 16th international conference on Text, Speech and Dialogue (TSD'13), Springer, Berlin/Heidelberg. pp. 153–160.

Kouloumpis, E., Wilson, T., Moore, J., 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.

Laboreiro, G., Sarmento, L., Teixeira, J., Oliveira, E., 2010. Tokenizing micro-blogging messages using a text classification approach, in: Proceedings of the fourth workshop on Analytics for noisy unstructured text data, ACM, New York, NY, USA. pp. 81–88.

Liu, B., Zhang, L., 2012. A survey of opinion mining and sentiment analysis, in: Mining Text Data. Springer, pp. 415–463.

Liu, H., Sun, J., Liu, L., Zhang, H., 2009. Feature selection with dynamic mutual information. Pattern Recogn. 42, 1330–1339.

López, R., Tejada, J., Thelwall, M., 2012. Spanish sentistrength as a tool for opinion mining peruvian facebook and twitter, in: Artificial Intelligence Driven Solutions to Business and Engineering Problems. ITHEA, Sofia, Bulgaria, pp. 82–85.

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.

Martineau, J., Finin, T., 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.

Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M.T., Ureña López, L.A., 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, Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 3–10.

Pak, A., Paroubek, P., 2010. Twitter as a corpus for sentiment analysis and opinion mining, in: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (Eds.), Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, European Language Resources Association.

Paltoglou, G., Thelwall, M., 2010. A study of information retrieval weighting schemes for sentiment analysis, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 1386–1395.

Pang, B., Lee, L., Vaithyanathan, S., 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, Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 79–86.

Patočka, M., 2013. Machine Learning for Sentiment Analysis. Master's thesis. University of West Bohemia. Plzen, Czech Republic. [In Czech].

Pustejovsky, J., Stubbs, A., 2013. Natural Language Annotation for Machine Learning. O'Reilly Media, Sebastopol, CA 95472.

Sharma, A., Dey, S., 2012. A comparative study of feature selection and machine learning techniques for sentiment analysis, in: Proceedings of the 2012 ACM Research in Applied Computation Symposium, ACM, New York, NY, USA. pp. 1–7.

Steinberger, J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M.A., Lenkova, P., Steinberger, R., Tanev, H., Vázquez, S., Zavarella, V., 2012. Creating sentiment dictionaries via triangulation. Decision Support Systems 53, 689—694.

Steinberger, J., Lenkova, P., Kabadjov, M.A., Steinberger, R., der Goot, E.V., 2011. Multilingual entity-centered sentiment analysis evaluated by parallel corpora, in: Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing, pp. 770–775.

Stepanov, E., Riccardi, G., 2011. Detecting general opinions from customer surveys, in: Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, pp. 115–122.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M., 2011. Lexicon-based methods for sentiment analysis. Computational Linguistics 37, 267–307.

Tsytsarau, M., Palpanas, T., 2012. Survey on mining subjective data on the web. Data Mining and Knowledge Discovery 24, 478–514.

Uchyigit, G., 2012. Experimental evaluation of feature selection methods for text classification, in: Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on, pp. 1294–1298. doi:10.1109/FSKD.2012.6234191.

Veselovská, K., Jr., J.H., Šindlerová, J., 2012. Creating annotated resources for polarity classification in Czech, in: Proceedings of KONVENS 2012, ÖGAI. pp. 296–304. PATHOS 2012 workshop.

Wasikowski, M., Chen, X.w., 2010. Combating the small sample class imbalance problem using feature selection. IEEE Trans. on Knowl. and Data Eng. 22, 1388–1400.

Yu, L.C., Wu, J.L., Chang, P.C., Chu, H.S., 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.

Zhang, D., Si, L., Rego, V.J., 2012. Sentiment detection with auxiliary data. Information Retrieval 15, 373–390.
Zhang, K., Cheng, Y., Xie, Y., Honbo, D., Agrawal, A., Palsetia, D., Lee, K., keng Liao, W., Choudhary, A.N., 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, IEEE. pp. 129–136.
Zheng, Z., Wu, X., Srihari, R., 2004. Feature selection for text categorization on imbalanced data. SIGKDD Explor. Newsl. 6, 80–89. URL: http://doi.acm.org/10.1145/1007730.1007741, doi:10.1145/1007730.1007741.