Sarcasm Detection on Czech and English Twitter

Tomáš Ptáček[†‡], **Ivan Habernal**[†] **and Jun Hong**[‡]
[†] 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 habernal@kiv.zcu.cz

[‡] School of Electronics, Electrical Engineering and Computer Science,
Queen's University Belfast, Belfast BT7 1NN, UK
j.hong@qub.ac.uk

Abstract

This paper presents a machine learning approach to sarcasm detection on Twitter in two languages – English and Czech. Although there has been some research in sarcasm detection in languages other than English (e.g., Dutch, Italian, and Brazilian Portuguese), our work is the first attempt at sarcasm detection in the Czech language. We created a large Czech Twitter corpus consisting of 7,000 manually-labeled tweets and provide it to the community. We evaluate two classifiers with various combinations of features on both the Czech and English datasets. Furthermore, we tackle the issues of rich Czech morphology by examining different preprocessing techniques. Experiments show that our language-independent approach significantly outperforms adapted state-of-the-art methods in English (F-measure 0.947) and also represents a strong baseline for further research in Czech (F-measure 0.582).

1 Introduction

Sentiment analysis on social media has been one of the most targeted research topics in NLP in the past decade, as shown in several recent surveys (Liu and Zhang, 2012; Tsytsarau and Palpanas, 2012). Since the goal of sentiment analysis is to automatically detect the polarity of a document, misinterpreting irony and sarcasm represents a big challenge (Davidov et al., 2010).

As there is only a weak boundary in meaning between irony, sarcasm and satire (Reyes et al., 2012), we will use only the term sarcasm in this paper. Bosco et al. (2013) claim that "even if there is no agreement on a formal definition of irony, psychological experiments have delivered evidence that humans can reliably identify ironic text utterances from an early age in life." We have thus decided to rely on the ability of our human annotators to manually label sarcastic tweets to train our classifiers. Sarcasm generally reverses the polarity of an utterance from positive or negative into its opposite, which deteriorates the results of a given NLP task. Therefore, correct identification of sarcasm can improve the performance.

The issue of automatic sarcasm detection has been addressed mostly in English, although there has been some research in other languages, such as Dutch (Liebrecht et al., 2013), Italian (Bosco et al., 2013), or Brazilian Portuguese (Vanin et al., 2013). To the best of our knowledge, no research has been conducted in Czech or other Slavic languages. These languages are challenging for many NLP tasks because of their rich morphology and syntax. This has motivated us to focus our current research on both English and Czech.

Majority of the existing state-of-the-art techniques are language dependent, which rely on language-specific lexical resources. Since no such resources are available for Czech, we adapt some language-independent methods and also apply various preprocessing steps for sentiment analysis proposed by Habernal et al. (2013).

This paper focuses on document-level sarcasm detection on Czech and English Twitter datasets using supervised machine learning methods. The Czech dataset consists of 7,000 manually labeled tweets,

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 213–223, Dublin, Ireland, August 23-29 2014.

the English dataset consists of a balanced distribution and an imbalanced distribution, each containing 100,000 tweets, where hashtag #sarcasm was used as an indicator of sarcastic tweets. We provide both datasets under Creative Commons BY-NC-SA licence[1] at http://liks.fav.zcu.cz/sarcasm/.

Our research questions were the following: (1) To what extent can the language-independent approach compete with methods based on lexical language-dependent resources? (2) Is it possible to reach good agreement on annotating sarcasm and what typical text properties on Twitter are important for sarcasm detection? (3) What is the best preprocessing pipeline that can boost performance on highly-flective Czech language and what types of features and classifiers yield the best results?

The rest of this article is organized as follows. Section 2 describes the related work. In section 3, we outline our approach to sarcasm detection and describe the selection of features in our approach. Section 4 thoroughly describes the datasets and the annotation process. Section 5 describes and discusses the experimental results. Finally we conclude in Section 6.

2 Related Work

Experiments with semi-supervised sarcasm identification on a Twitter dataset (5.9 million tweets) and on 66,000 product reviews from Amazon were conducted in (Davidov et al., 2010) and (Tsur et al., 2010). They used 5-fold cross validation on their kNN-like classifier and obtained an F-measure of 0.83 on the product reviews dataset and 0.55 on the Twitter dataset. For acquiring the Twitter dataset they used hashtag #sarcasm as an indicator of sarcastic tweets. They further created a balanced evaluation set of 180 tweets using 15 human annotators via Amazon Mechanical Turk[2] and achieved an inter-annotator agreement 0.41 (Fleiss' κ).

González-Ibáñez et al. (2011) experimented with Twitter data divided into three categories (sarcastic, positive sentiment and negative sentiment), each containing 900 tweets. They used the #sarcasm and #sarcastic hashtags to identify sarcastic tweets. They used two classifiers – support vector machine (SVM) with sequential minimal optimization (SMO) and logistic regression. They tried various combinations of unigrams, dictionary-based features and pragmatic factors (positive and negative emoticons and user references), achieving the best result (accuracy 0.65) for sarcastic and non-sarcastic classification with the combination of SVM with SMO and unigrams. They employed 3 human judges to annotate 180 tweets (90 sarcastic and 90 non-sarcastic). The human judges achieved Fleiss' $\kappa = 0.586$, demonstrating the difficulty of sarcasm classification. Another experiment included 50 sarcastic and 50 non-sarcastic (25 positive, 25 negative) tweets with emoticons annotated by two judges. The automatic classification and human judges achieved the accuracy of 0.71 and 0.89 respectively. The inter-annotator agreement (Cohen's κ) was 0.74.

Reyes et al. (2012) proposed features to capture properties of a figurative language such as ambiguity, polarity, unexpectedness and emotional scenarios. Their corpus consists of five categories (humor, irony, politics, technology and general), each containing 10,000 tweets. The best result in the classification of irony and general tweets was F-measure 0.65.

In (Reyes et al., 2013) they explored the representativeness and relevance of conceptual features (signatures, unexpectedness, style and emotional scenarios). These features include punctuation marks, emoticons, quotes, capitalized words, lexicon-based features, character n-grams, skip-grams, (Guthrie et al., 2006), and polarity skip-grams. Their corpus consists of four categories (irony, humor, education and politics), each containing 10,000 tweets. Their evaluation was performed on two distributional scenarios, balanced distribution and imbalanced distribution (25% ironic tweets and 75% tweets from all three non-ironic categories) using the Naive Bayes and decision trees algorithms from the Weka toolkit (Witten and Frank, 2005). The classification by the decision trees achieved an F-measure of 0.72 on the balanced distribution and an F-measure of 0.53 on the imbalanced distribution.

The work of Riloff et al. (2013) identifies one type of sarcasm: contrast between a positive sentiment and negative situation. They used a bootstrapping algorithm to acquire lists of positive sentiment phrases

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

[2] http://www.mturk.com

and negative situation phrases from sarcastic tweets. They proposed a method which classifies tweets as sarcastic if it contains a positive predicative that precedes a negative situation phrase in close proximity. Their evaluation on a human-annotated dataset[3] of 3000 tweets (23% sarcastic) was done using the SVM classifier with unigrams and bigrams as features, achieving an F-measure of 0.48. The hybrid approach that combines the results of the SVM classifier and their contrast method achieved an F-measure of 0.51.

Sarcasm and nastiness classification in online dialogues was also explored in (Lukin and Walker, 2013) using bootstrapping, syntactic patterns and a high precision classifier. They achieved an F-measure of 0.57 on their sarcasm dataset.

3 Our Approach

This paper presents the first attempt at sarcasm detection in the Czech language, in which we focus on supervised machine learning approaches and evaluate their performance. We selected various n-grams, including unigrams, bigrams, trigrams with frequency greater than three (Liebrecht et al., 2013), and a set of language-independent features, including punctuation marks, emoticons, quotes, capitalized words, character n-grams and skip-grams (Reyes et al., 2013) as our baselines.

3.1 Classification

Our evaluation was performed using the Maximum Entropy (MaxEnt) and Support Vector Machine (SVM) classifiers. We used *Brainy* – a Java framework for machine learning (Konkol, 2014) – with default settings (the linear kernel for SVM). All experiments were conducted in the 5-fold cross validation manner similar to (Davidov et al., 2010; González-Ibáñez et al., 2011). Our motivation to test multiple classifiers stemmed also from related works which mostly test more than one classifier. On the other hand, the choice between state-of-the-art linear classifiers might not be much of importance, as the most important is the feature engineering.

3.2 Features

For our evaluation we used the most promising language-independent features from the related work and POS related features. Feature sets used in our evaluation are described in Table 1.

Group	Features	Description
N-gram	Character n-gram	We used character n-gram features (Blamey et al., 2012). We set the minimum occurrence of a particular character n-gram to either 5 or 50, in order to prune the feature space. Our character feature set contains 3-grams to 6-grams.
	N-gram	We used word unigrams, bigrams and trigrams as binary features. The feature space is pruned by the minimum n-gram occurrence set to 3 (Liebrecht et al., 2013).
	Skip-bigram	Instead of using sequences of adjacent words (n-grams) we used skip-grams (Guthrie et al., 2006), which skip over arbitrary gaps. Reyes et al. (2013) consider skip-bigrams with 2 or 3 word skips and remove skip-grams with a frequency ≤ 20.
Pattern	Pattern	Patterns composed of high frequency words (HFWs)[4] and content words (CWs)[5] used by (Davidov et al., 2010). Pattern must contain at least one high frequency word. The patterns contain 2-6 HFWs and 1-6 CWs. We set the minimum occurrence of a particular pattern to 5.
	Word-shape pattern	We tried to improve pattern features by using word-shape classes for content words. We assign words into one of 24 classes[6] similar to the function specified in (Bikel et al., 1997).
POS	POS characteristics	We implemented various POS features, 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 obtained from POS tags.

[3]They used three annotators. Each annotator was given the same 100 tweets with the sarcasm hashtag and 100 tweets without the sarcasm hashtag (the hashtags were removed). On these tweets the pairwise inter-annotator scores were computed (Cohen's Kappa $\kappa_1 = 0.80, \kappa_2 = 0.81$ and $\kappa_3 = 0.82$). Then each annotator labeled additional 1000 tweets.

[4]A word whose corpus frequency is more than 1000 words per million plus all punctuation characters.

[5]A word whose corpus frequency is less than 1000 words per million.

[6]We use edu.stanford.nlp.process.WordShapeClassifier with the WORDSHAPECHRIS1 setting.

	POS word-shape	Unigram feature consisting of POS and word-shape (see Word-shape pattern). The feature space is pruned by the minimum occurrence set to 5.
	POS n-gram	Direct use of POS n-grams has not shown any significant improvement in sentiment analysis but it may improve the results of sarcasm detection. We experimented with 3-grams to 6-grams with the minimum n-gram occurrence set to 5.
Others	Emoticons	We used two lists of positive and negative emoticons (Montejo-Ráez et al., 2012). The feature captures the number of occurrences of each class of emoticons within the text.
	Punctuation-based	We adapted punctuation-based features proposed by (Davidov et al., 2010). This feature set consists of number of words, exclamation marks, question marks, quotation marks and capitalized words normalized by dividing them by the maximal observed value multiplied by the averaged maximal value of the other feature groups.
	Pointedness	Pointedness was used by (Reyes et al., 2013) to distinguish irony. It focuses on explicit marks which should reflect a sharp distinction in the information that is transmitted. The presence of punctuation marks, emoticons, quotes and capitalized words has been considered.
	Extended Pointedness	This feature captures the number of occurrences of punctuation marks and emoticons as well as the number of words, exclamation marks, question marks, quotation marks and capitalized words normalized by maximal observed value.
	Word-case	We implemented various word-case features that include, e.g., the number of upper cased words, number of words with first letter capital normalized by number of words and number of upper cased characters normalized by number of words.

Table 1: Descriptions of used feature sets.

4 Evaluation Datasets

We collected datasets using *Twitter Search API* and *Java Language Detector*[7]. We collected 140,000 Czech and 780,000 English tweets, respectively. Due to lack of support for the Czech language on Twitter, we used the *Twitter Search API* parameter *geocode* to acquire tweets posted near Prague. For the English dataset we also collected tweets with the #sarcasm hashtag. Czech users generally don't use the sarcasm ("`#sarkasmus`") or irony ("`#ironie`") hashtag variants[8] thus we had to annotate the Czech dataset manually. The final label distribution in datasets is shown in Table 4.

4.1 Filtering and Normalization

All user, URL and hashtag references in tweets have been replaced by "user", "link" and "hashtag" respectively. We also removed all tweets starting with "RT" because they refer to previous tweets and tweets containing just combinations of user, link, "RT" and hashtags without any additional words.

Tokenization of tweets requires proper handling of emoticons and other special character sequences typical on Twitter. The *Ark-tweet-nlp tool* (Gimpel et al., 2011) offers precisely that and although it was developed and tested in English, it yields satisfactory results in Czech as well.

Czech is a highly flective language and uses a lot of diacritics. However some Czech users type only the unaccented characters.[9] Preliminary experiments showed that removing diacritics yields better results, thus we removed diacritics from all tweets.

4.2 Czech Dataset Annotation

Firstly we conducted an experiment to determine whether to annotate the original data or the normalized data. We selected two sample sets of 50 tweets containing Czech sarcasm (`#sarkasmus`) and irony (`#ironie`) hashtags and other tweets. One annotator obtained the original data while the other got the normalized data from the first sample set. We then tried to give both annotators the original data from the first sample set and finally we gave them both the normalized data from the second sample set. Table 2 shows the difficulty of sarcasm identification without the knowledge hidden in hashtags, user and links.

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

[8] We found only 10 tweets with sarcasm hashtag ("`#sarkasmus`") and 100 tweets with irony hashtag ("`#ironie`") in 140,000 collected tweets.

[9] Approximately 10% of collected tweets were without any diacritics.

Normalized	Normalized		
	Tag	**n**	**s**
	n	35	10
	s	0	5
Cohen's κ:		0.412	

Original	Normalized		
	Tag	**n**	**s**
	n	19	10
	s	5	16
Cohen's κ:		0.404	

Original	Original		
	Tag	**n**	**s**
	n	25	4
	s	3	18
Cohen's κ:		0.715	

Table 2: Confusion matrices and annotation agreement (Cohen's κ) between two annotators using original or normalized data.

"Basic" pipe	**Pipe 2**	**Pipe 3**
Tokenizing: ArkTweetNLP		
POS tagging: PDT		
–	**Stem:** no (Sn) / light (Sl) / HPS (Sh)	
–	**Stopwords removal**	
–	–	**Phonetic:** eSpeak (Pe)

Table 3: The preprocessing pipes for Czech (top-down). Combinations of methods are denoted using the appropriate labels, e.g. "Sn" means *1. tokenizing, 2. POS-tagging, 3. no stemming and 4. removing stopwords*. eSpeak stands for a phonetic transcription to International Phonetic Alphabet, which should reduce the effects of grammar mistakes and misspellings.

The most promising results come from the annotation of the original data, thus the rest of the data are annotated in this manner.

We randomly selected 7,000 tweets from the collected data for annotation. The annotators were given just simple instructions without an explicit sarcasm definition (see Section 1): *"A tweet is considered sarcastic when its content is intended ironically / sarcastically without anticipating further information. Offensive utterances, jokes and ironic situations are not considered ironic / sarcastic."*

The complete dataset of 7,000 tweets was independently annotated by two annotators. The inter-annotator agreement (Cohen's κ) between the two annotators is 0.54. They disagreed on 403 tweets. To resolve these conflicts we used a third annotator.

The third annotator has been instructed the same way as the other two. The final κ agreement was measured between the first two annotators, thus it was not affected by the third annotator. Kappa agreements measured on the conflicted states (403 tweets) were 0.4 (annotator 1 vs. annotator 3) and 0.6 (annotator 2 vs. annotator 3).

Preprocessing

Preprocessing steps for handling social media texts in Czech were explored in (Habernal et al., 2013). The preprocessing diagram and its variants is depicted in Table 3. Overall, there are various possible preprocessing "pipe" configurations including "Basic" pipeline consisting of tokenizing and POS-tagging only. We adapted all their preprocessing pipelines. However, as the number of combinations would be too large, we report only the settings with better performance.

4.3 English Dataset

We collected 780,000 (130,000 sarcastic and 650,000 non-sarcastic) tweets in English. The `#sarcasm` hashtag was used as an indicator of sarcastic tweets. From this corpus we created two distributional scenarios based on the work of (Reyes et al., 2013). Refer to Table 4 for the final statistics of the dataset. Part of speech tagging was done using the *Ark-tweet-nlp tool* (Gimpel et al., 2011).

5 Results

For each preprocessing pipeline (refer to table 3) we assembled various sets of features and employed two classifiers. Accuracy (micro F-measure) tends to prefer performance on dominant classes in highly

Dataset \ Tweets	Sarcastic	Non-sarcastic
Czech	325	6,675
English Balanced	50,000	50,000
English Imbalanced	25,000	75,000

Table 4: The tweet distributions in datasets.

Feature Set \ Pipeline	Basic	Sh	ShPe	Sl	SlPe	Sn	SnPe
Baseline 1 (B1): n-gram	54.8	55.3	55.2	55.0	55.0	54.4	55.3
B1 + pattern	55.1	54.4	54.7	55.1	54.8	54.2	54.5
B1 + word-shape pattern	54.6	54.8	55.2	54.4	55.0	54.8	55.1
B1 + punctuation-based	54.7	48.8	48.8	48.8	48.8	53.8	55.5
B1 + pointedness	55.0	54.7	54.7	55.0	55.9	54.8	54.9
B1 + extended pointedness	54.5	48.8	48.8	48.8	48.8	54.7	54.6
B1 + POS n-gram	53.4	54.1	54.2	55.3	55.1	54.2	53.9
B1 + POS word-shape	55.0	55.6	55.2	54.8	54.6	55.8	54.4
B1 + skip-bigram	54.2	54.8	54.2	54.7	**56.0**	54.6	54.4
B1 + POS characteristics + emoticons	55.5	54.7	55.6	55.2	55.4	55.2	53.9
B1 + POS characteristics + emoticons + word-case	53.8	**56.4**	55.5	54.6	55.3	55.9	55.3
Character n-gram (3-6, min. occurrence > 5)	53.0	52.7	53.2	53.9	54.7	52.0	53.2
Baseline 2 (B2)	55.0	55.2	55.4	**56.8**	**56.2**	54.7	54.0
B2 + FS1	52.3	48.8	48.8	48.8	48.8	52.0	52.9
B2 + FS1 + FS2	53.0	48.8	48.8	48.8	48.8	52.2	53.6
B2 + pattern	55.3	55.4	55.7	**56.9**	**56.6**	54.4	53.6
B2 + POS word-shape	55.5	55.8	55.4	**57.0**	**56.3**	55.3	54.7
B2 + POS characteristics + emoticons + word-case	**56.1**	55.7	55.7	**56.9**	**56.1**	55.0	54.3

Table 5: Results on the Czech dataset with the MaxEnt classifier. Macro F-measure, 95% confidence interval $\approx \pm 1.2$. Best results are in bold. **B2**: character n-gram (3-5, min. occurrence > 50) + skip-bigram + pointedness; **FS1**: character n-gram (3-6, min. occurrence > 5) + extended pointedness; **FS2**: POS word-shape + pattern + POS characteristics + emoticons + word-case.

unbalanced datasets (Manning et al., 2008), thus we chose macro F-measure as the evaluation metric (Forman and Scholz, 2010), as it allows us to compare classification results on different datasets. For statistical significance testing, we report confidence intervals at α 0.05. Another applicable methods would be, i.e., two-matched-samples t Test or McNemar's test (Japkowicz and Shah, 2011).

5.1 Czech

Tables 5 and 6 show the results on the Czech dataset. The best result (F-measure 0.582) was achieved by the SVM classifier and a feature set enriched with patterns, utilizing stopwords removal and phonetic transcription in the preprocessing step.

The importance of the appropriate preprocessing techniques for Czech is evident from the improvement of results for various feature sets, e.g., the best result for "Basic" pipeline (see line "B2 + pattern"). Both baselines show improvement on most preprocessing pipelines. The most significant difference is visible on the second baseline with the MaxEnt classifier and the "Sl" pipeline where the F-measure is 0.018 higher than the "Basic" pipeline with no additional preprocessing. The n-gram baseline was significantly outperformed by the SVM classifier with feature sets "B1 + POS characteristics + Emoticons + Word-case" and "B1 + extended pointedness" on the "SnPe" pipeline.

Error Analysis

To get a better understanding of the limitations of our approach, we inspected 100 random tweets from the Czech dataset, which were wrongly classified by the SVM classifier with the best feature combination.

Feature Set \ Pipeline	Basic	Sh	ShPe	Sl	SlPe	Sn	SnPe
Baseline 1 (B1): n-gram	55.8	54.6	54.5	54.6	55.5	56.0	53.9
B1 + pattern	55.6	54.0	54.3	54.6	55.7	55.4	55.6
B1 + word-shape pattern	54.9	55.0	53.8	55.2	55.1	55.4	55.3
B1 + punctuation-based	55.8	48.8	48.8	48.8	48.8	55.7	53.7
B1 + pointedness	55.9	54.5	53.1	54.6	54.3	55.4	54.6
B1 + extended pointedness	56.5	48.8	48.8	48.8	48.8	55.8	56.9
B1 + POS n-gram	54.0	54.1	54.0	54.7	53.4	54.5	53.9
B1 + POS word-shape	55.2	56.4	55.9	55.1	56.0	56.1	55.0
B1 + skip-bigram	55.9	55.3	54.8	55.4	55.0	56.1	55.2
B1 + POS characteristics + emoticons	55.9	54.5	54.1	54.6	54.2	56.7	55.8
B1 + POS characteristics + emoticons + word-case	55.6	54.5	54.3	55.1	55.5	56.3	56.4
Character n-gram (3-6, min. occurrence > 5)	54.6	53.6	53.3	55.2	53.6	53.4	54.9
Baseline 2 (B2)	55.9	56.4	56.3	**57.0**	56.2	**57.1**	55.8
B2 + FS1	52.2	48.8	48.8	48.8	48.8	53.1	52.7
B2 + FS1 + FS2	54.0	48.8	48.8	48.8	48.8	54.4	54.3
B2 + pattern	56.8	**57.0**	56.7	56.5	**57.5**	**57.1**	**58.2**
B2 + POS word-shape	56.5	56.3	**57.2**	56.4	56.1	56.3	**57.8**
B2 + POS characteristics + emoticons + word-case	56.2	55.7	55.8	56.0	56.0	**57.0**	56.0

Table 6: Results on the Czech dataset with the SVM classifier. Macro F-measure, 95% confidence interval $\approx \pm 1.2$. Best results are in bold. **B2**: character n-gram (3-5, min. occurrence > 50) + skip-bigram + pointedness; **FS1**: character n-gram (3-6, min. occurrence > 5) + extended pointedness; **FS2**: POS word-shape + pattern + POS characteristics + emoticons + word-case.

We found 48 false positives and 52 false negatives. The annotators disagreed upon 10% of these tweets.

Non-sarcastic tweets were often about news, reviews, general information and user status updates. In most of the difficult cases of true negatives, the tweet contains a question, insult, opinion or wordplay.

Understanding sarcasm in some tweets was often bound with broader common knowledge (e.g., about news or celebrities), the context known only to the author or authors opinion. Another difficulty poses subtle or sophisticated expression of sarcasm such as *"I'm not sure whether you didn't overdo a bit the first part of the renovation - the demolition. :)"*[10] or *"Conservatism, once something is in the school rules, it must be followed, forever, otherwise anarchy will break out and traditional values will die."*[11]

5.2 English

The results on both balanced and imbalanced English datasets are presented in Table 7. In most cases the MaxEnt classifier significantly outperforms the SVM classifier. The combination of majority of features ("B2 + FS1 + FS2") with the MaxEnt classifier yields the best results for both balanced and imbalanced dataset distributions. This suggests that these features are coherent. While no single feature captures the essence of sarcasm, all features together provide useful linguistic information for detecting sarcasm at textual level.

Balanced distribution Both baselines were surpassed by various combinations of feature sets with the MaxEnt classifier, although in some cases very narrowly ("B1 + punctuation-based" and "B1 + pointedness" feature sets). Although the SVM classifier has slightly worse results, it still performs reasonably, and we even recorded significant improvement over the baseline for "B1 + POS word-shape". The best results were achieved using the MaxEnt classifier with "B2 + FS1 + FS2" (F-measure 0.947) and "B1 + word-shape pattern " (F-measure 0.943) feature sets.

[10] "Jestli jste tu první část rekonstrukce - demolici - trochu nepřehnali . :)"

[11] "Konzervatismus , když je to jednou ve školním řádu , tak se to musí dodržovat , a to navždy , jinak vypukne anarchie a tradiční hodnoty zemřou ."

Dataset	Balanced				Imbalanced			
Classifier	MaxEnt		SVM		MaxEnt		SVM	
Feature set \ Results	Fm	CI	Fm	CI	Fm	CI	Fm	CI
Baseline 1 (B1): n-gram	**93.28**	0.16	92.86	0.16	90.76	0.18	90.44	0.18
B1 + pattern	**94.25**	0.14	**93.13**	0.16	91.86	0.17	90.22	0.18
B1 + word-shape pattern	**94.33**	0.14	**93.17**	0.16	**92.01**	0.17	90.35	0.18
B1 + punctuation-based	**93.32**	0.15	92.84	0.16	90.72	0.18	90.43	0.18
B1 + pointedness	**93.29**	0.16	92.99	0.16	91.00	0.18	90.07	0.19
B1 + extended pointedness	**93.68**	0.15	92.61	0.16	91.07	0.18	89.89	0.19
B1 + POS n-gram	**93.66**	0.15	92.64	0.16	91.20	0.18	89.85	0.19
B1 + POS word-shape	**93.96**	0.15	**93.19**	0.16	91.41	0.17	90.51	0.18
B1 + skip-bigram	**93.63**	0.15	**93.17**	0.16	90.99	0.18	90.48	0.18
B1 + POS characteristics + emoticons	**93.97**	0.15	91.66	0.17	91.69	0.17	89.39	0.19
B1 + POS characteristics + emoticons + word-case	**93.96**	0.15	91.54	0.17	91.61	0.17	88.89	0.19
Character n-gram: (3-6, min. occurrence > 5)	**93.01**	0.16	91.73	0.17	90.36	0.18	88.81	0.20
Baseline 2 (B2)	92.81	0.16	91.67	0.17	90.65	0.18	88.70	0.20
B2 + FS1	**93.82**	0.15	91.56	0.17	91.21	0.18	88.73	0.20
B2 + FS1 + FS2	**94.66**	0.14	91.39	0.17	**92.37**	0.16	88.62	0.20
B2 + pattern	**93.60**	0.15	91.66	0.17	90.86	0.18	88.82	0.20
B2 + POS word-shape	**93.20**	0.16	91.65	0.17	90.82	0.18	88.74	0.20
B2 + POS characteristics + emoticons + word-case	**93.21**	0.16	91.07	0.18	89.98	0.19	88.40	0.20

Table 7: Results on the English dataset with the MaxEnt and SVM classifiers. Macro F-measure (Fm) and 95% confidence interval (CI) are in %. Best results are in bold. **B2**: character n-gram (3-5, min. occurrence > 50) + skip-bigram + pointedness; **FS1**: character n-gram (3-6, min. occurrence > 5) + extended pointedness; **FS2**: POS word-shape + pattern + POS characteristics + emoticons + word-case.

Imbalanced distribution However, data in the real world do not necessarily resemble the balanced distribution. Therefore we have also performed the evaluation on an imbalanced distribution. The MaxEnt classifier clearly achieves the best results. This experiment indicates that combinations of features "B2 + FS1 + FS2" (F-measure 0.924) and "B1, word-shape pattern" (F-measure 0.920) yields the best results for both balanced and imbalanced dataset distribution.

5.3 Discussion

To explain the huge difference in the performance between English and Czech, we conducted an additional experiment in English. We sampled the "big-data" English corpus (100k Tweets) to obtain the same distribution as on the "small-data" Czech corpus (325 sarcastic and 6,675 non-sarcastic Tweets). Feature combination "B2 + FS1 + FS2" achieves an F-measure of 0.734 ± 0.01 (MaxEnt classifier) and 0.729 ± 0.01 (SVM). This performance drop shows that the amount of training data plays a key role (≈ 0.92 on "big-data" vs. ≈ 0.73 on "small-data"). However, these results are still significantly better than in Czech (≈ 0.58). This demonstrates that Czech is a challenging language in sarcasm detection, as in other NLP tasks.

In addition, we also experimented with the Naive Bayes classifier and with delta TF-IDF feature variants (Martineau and Finin, 2009; Paltoglou and Thelwall, 2010) in both languages. However, the performance was not satisfactory in comparison with the reported results.

6 Conclusions

We investigated supervised machine learning methods for sarcasm detection on Twitter. As a pilot study for sarcasm detection in the Czech language, we provide a large human-annotated Czech Twitter dataset containing 7,000 tweets with inter-annotator agreement $\kappa = 0.54$. The novel contributions of our work include the extensive evaluation of two classifiers with various combinations of feature sets on both the Czech and English datasets as well as a comparison of different preprocessing techniques for the

Czech dataset. Our approaches significantly outperformed both baselines adapted from related work[12] in English and achieved F-measure of 0.947 and 0.924 on the balanced and imbalanced datasets, respectively.[13] The best result on the Czech dataset was achieved by the SVM classifier with the feature set enriched with patterns yielding F-measure 0.582. The whole project is available to the community under GPL license at `http://liks.fav.zcu.cz/sarcasm/`. We believe that our findings will contribute to the research outside the mainstream languages and may be applied to sarcasm detection in other Slavic languages, such as Slovak or Polish.

6.1 Future work

We approached the problem mainly from the data-driven perspective (annotation, feature engineering, error analysis). However, we feel that elaborating deep linguistic insights would be helpful to better understand the phenomena of sarcasm on social media (Averbeck, 2013; Averbeck and Hample, 2008; Ivanko et al., 2004; Jorgensen, 1996).

There are also possible extensions to the lexical/morphological features – either in the direction of semi-supervised learning and adding for example features based on latent semantics, topic models, or graphical models popular in the sentiment analysis field (Habernal and Brychcín, 2013; Brychcín and Habernal, 2013), or the direction of deeper linguistic processing in terms of, e.g., syntax/dependecy parsing (but this has limitation given the nature of Twitter data as well as unavailability of such tools for Czech). These deserve further investigation and are planned in future work.

Acknowledgements

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 greatly appreciated. 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.

References

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

Joshua M Averbeck and Dale Hample. 2008. Ironic message production: How and why we produce ironic messages. *Communication Monographs*, 75(4):396–410.

Joshua M Averbeck. 2013. Comparisons of ironic and sarcastic arguments in terms of appropriateness and effectiveness in personal relationships. *Argumentation & Advocacy*, 50(1).

Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, pages 194–201. Association for Computational Linguistics.

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.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28(2):55–63.

Tomáš Brychcín and Ivan Habernal. 2013. Unsupervised improving of sentiment analysis using global target context. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 122–128, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

[12] Word unigrams, bigrams, trigrams (Liebrecht et al., 2013) and a set of language-independent features (punctuation marks, emoticons, quotes, capitalized words, character n-grams and skip-grams.) (Reyes et al., 2013)

[13] Note that the best result (F-measure 0.715 on the balanced distribution and F-measure 0.533 on the imbalanced distribution) from the related work was achieved by (Reyes et al., 2013) using decision trees classifier.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

George Forman and Martin Scholz. 2010. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *SIGKDD Explor. Newsl.*, 12(1):49–57, 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.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 581–586, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. 2006. A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, pages 1–4.

Ivan Habernal and Tomáš Brychcín. 2013. Semantic spaces for sentiment analysis. In Ivan Habernal and Václav Matoušek, editors, *Text, Speech, and Dialogue*, volume 8082 of *Lecture Notes in Computer Science*, pages 484–491. Springer Berlin Heidelberg.

Ivan Habernal, Tomáš Ptáček, and Josef Steinberger. 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*, pages 65–74, Atlanta, Georgia, June. Association for Computational Linguistics.

Stacey L Ivanko, Penny M Pexman, and Kara M Olineck. 2004. How sarcastic are you? individual differences and verbal irony. *Journal of language and social psychology*, 23(3):244–271.

Nathalie Japkowicz and Mohak Shah. 2011. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press.

Julia Jorgensen. 1996. The functions of sarcastic irony in speech. *Journal of Pragmatics*, 26(5):613 – 634.

Michal Konkol. 2014. Brainy: A machine learning library. In Leszek Rutkowski, Marcin Korytkowski, Rafal Scherer, Ryszard Tadeusiewicz, Lotfi Zadeh, and Jacek Zurada, editors, *Artificial Intelligence and Soft Computing*, volume 8468 of *Lecture Notes in Computer Science*, pages 490–499. Springer International Publishing.

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.

Christine Liebrecht, Florian Kunneman, and Antal Van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, Atlanta, Georgia, June. Association for Computational Linguistics.

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

Stephanie Lukin and Marilyn Walker. 2013. Really? Well. Apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 30–40, Atlanta, Georgia, June. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 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.

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.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data Knowl. Eng.*, 74:1–12, April.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239–268.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA, October. Association for Computational Linguistics.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. ICWSM - a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In William W. Cohen and Samuel Gosling, editors, *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*. The AAAI Press.

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

Aline A Vanin, Larissa A Freitas, Renata Vieira, and Marco Bochernitsan. 2013. Some clues on irony detection in tweets. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 635–636. International World Wide Web Conferences Steering Committee.

Ian H Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.