A Comparison of Approaches for Sentiment Classification on Lithuanian Internet Comments

Jurgita Kapočiūtė-Dzikienė
Kaunas University of Technology / K. Donelaičio 73,
LT-44249 Kaunas, Lithuania
jurgita.k.dz@gmail.com

Algis Krupavičius
Kaunas University of Technology / K. Donelaičio 73,
LT-44249 Kaunas, Lithuania
pvai@ktu.lt

Tomas Krilavičius
Baltic Institute of Advanced Technology / Saulėtekio 15,
LT-10224 Vilnius, Lithuania
t.krilavicius@bpti.lt

Abstract

Despite many methods that effectively solve sentiment classification task for such widely used languages as English, there is no clear answer which methods are the most suitable for the languages that are substantially different. In this paper we attempt to solve Internet comments sentiment classification task for Lithuanian, using two classification approaches – knowledge-based and supervised machine learning. We explore an influence of sentiment word dictionaries based on the different parts-of-speech (adjectives, adverbs, nouns, and verbs) for knowledge-based method; different feature types (bag-of-words, lemmas, word n-grams, character n-grams) for machine learning methods; and pre-processing techniques (emoticons replacement with sentiment words, diacritics replacement, etc.) for both approaches. Despite that supervised machine learning methods (Support Vector Machine and Naïve Bayes Multinomial) significantly outperform proposed knowledge-based method all obtained results are above baseline. The best accuracy 0.679 was achieved with Naïve Bayes Multinomial and token unigrams plus bigrams, when pre-processing involved diacritics replacement.

1 Introduction

An automatic extraction of opinions from a text has become an area of growing interest in the recent years. Due to the user-generated content available on the Internet companies can measure the feedback about their products or services; sociologists can look at people's reaction about public events; psychologists can study general mind-state of communities with regard to various issues; etc. Thus sentiment classification helps solving many various tasks, ranging from a very general to the very specific, requiring special solutions. Majority of tasks consider the content in general by focusing on the subjectivity vs. objectivity or semantic orientation (positive vs. negative) detection of reviews, tweets, blogs, or Internet comments. Others are solving very specific tasks, e.g. early threats detection (Bouma et al., 2012), prediction of user's potentiality to send out offensive content (Chen et al., 2012), etc.

But even adaptation to the task is not always effective due to the variations and complexity of the language. Sentiments are not always expressed explicitly, while for the meanings hidden in the context additional world knowledge is necessary. Moreover, sentiments may involve sarcasm and be interpreted differently in various domains and contexts. Despite all the mentioned difficulties, sentiment classification task is rather easy for us, humans, but manual analysis is time consuming and requires a lot of human-resources. Due to this fact automatic sentiment classifiers are often selected instead.

Various classification techniques effectively solve sentiment classification task for such widely used languages as English, but there is no clear answer which method is the most suitable for Lithuanian. Our focus is at finding classification approach yielding the best results on Lithuanian Internet comments by classifying them into positive, negative and neutral categories.

2 Related Work

Due to the complexity of sentiment classification task, there is a vast variety of methods trying to tackle this problem (for review see Pang and Lee (2008)).

All methods used to solve sentiment classification task fall into the three main categories: knowledge-based, machine learning and hybrid.

Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pages 2–11,
Sofia, Bulgaria, 8-9 August 2013. ©2010 Association for Computational Linguistics

In knowledge-based approaches sentiment is seen as the function of keywords (usually based on their count). Thus the main task is the construction of sentiment discriminatory-word lexicons with indicated class labels (positive or negative) and sometimes even with their intensiveness. Lexicons are constructed either manually (Taboada et al., 2011) or semi-automatically making use of such resources as WordNet (Hu and Liu, 2004); (Esuli and Sebastiani, 2006) or via word associations based on the heuristics evaluating word's occurrence alongside with the "seed" words in the text (Turney, 2002); (Turney and Littman, 2003).

Adjectives (or adjectival phrases) are considered as the most popular sentiment indicators, e.g. Benamara et al. (2007) claim that adjectives and adverbs (chosen based on the proposed adverb scoring technique) give much better results than adjectives alone; Taboada et al. (2011) show that such lexical items as nouns and verbs (not only adjectives and adverbs) can also carry important semantic polarity information.

Ding and Liu (2007) argue that semantic orientation is content dependent task and words alone are not sufficient sentiment indicators thus incorporate them into the set of linguistic rules used in classification; Choi and Cardie (2008) use heuristics based on the compositional semantics (considering the effect of interactions among the words) and achieve better results over the methods not incorporating it; Taboada et al. (2011) take into account valence shifters (intensifiers, downtoners, negation and irrealis markers) that influence the polarity of the neighboring words for English; Kuznetsova et al. (2013) – for Russian.

An alternative for the knowledge-based methods is machine learning that in turn can be grouped into supervised and clustering techniques. Clustering is rarely used due to the low accuracy, but the drawback of supervised machine learning (that we will further focus on) is that for model creation a training dataset (with manually pre-assigned sentiment class labels) is required.

The main issue for supervised machine learning techniques is proper selection of features. Nevertheless, the most basic approach remains bag-of-words interpretation. Pang et al. (2002) show that bag-of-words beat other feature types (based on token bigrams, parts-of-speech information and word position in the text) with Support Vector Machine (SVM) method. But on the contrary, Dave et al. (2003) report that token n-grams (up to trigrams) can improve the performance compared with simple unigrams; Cui et al. (2006) with higher order token n-grams (n = 3, 4, 5, 6) and Passive Aggressive classifier outperform unigrams and bigrams; Pak and Parubek (2011) with token bigrams and Naïve Bayes Multinomial method outperform both token unigrams and trigrams.

Dave et al. (2003) also report that stemming improves accuracy compared with the bag-of-words baseline, but other linguistic features (negation, collocations of words, etc.) on the contrary – hurt the performance. Raaijmakers and Kraaij (2008) use document-level character n-grams (n = 2, 3, 4, 5, 6) with SVM (geodesic kernel); Hartmann et al. (2011) claim that document-level character n-grams used, namely, with Naïve Bayes method are even better choice than token n-grams (because the probability of finding character n-gram is much higher and the relations between consecutive words are still considered).

Hybrid approaches combine both knowledge-based and machine learning methods thus achieving superior performance. As it is demonstrated by Mullen and Collier (2004) using SVM and combined token unigram features with those based on favorability measures (for phrases, adjectives and even knowledge of topic).

Sentiment classification results can be influenced by pre-processing as well. E.g. Kennedy and Inkpen (2006) claim that valence shifters and Mukherjee and Bhattacharyya (2012) show that discourse information incorporated into bag-of-words improve classification accuracy both for knowledge-based and SVM methods. But often pre-processing techniques (such as emoticons replacement, negation treatment and stop words removal) are selected without any considerations (e.g. see in (Pak and Paroubek, 2011)).

Both knowledge-based and supervised machine learning methods are domain-dependent (when classifier trained in one domain can barely beat the baseline in the other) and, moreover, domain-sensitive. E.g. Aue and Gamon (2005) with Naïve Bayes and SVM classifiers show that different types of features work better across different domains; therefore usually methods are built for the specific selected domain. Sometimes domain-dependent problem is circumvented by extracting related content with manually created rules (Wang et al., 2012) or via machine learning: i.e. by performing topicality classification at the first step and sentiment afterwards (Hurst and Nigam, 2004). Read and Carroll (2009) solve domain-depended problem by using special methodology to build the classifiers that are robust across the different domains.

Hence sentiment classification is domain and task dependent problem. Moreover, the performance of selected method can also depend on the language. E.g. Boiy and Moens (2009) demonstrate that the best accuracy with token unigrams (augmented with linguistics features) is obtained using Naïve Bayes Multinomial for English, SVM for Dutch and Maximum Entropy for French language. Besides, some solutions are proposed for multilingual texts as well, e.g. Cheng and Zhulyn (2012) show that generalized bigram model (especially suitable for the languages with a flexible word order) using Naïve Bayes and logistic regression classifiers can achieve high accuracy on different Germanic, Roman and East Asian languages.

We cannot provide any example of experiments based on sentiment classification for Lithuanian. Consequentially, this paper is the first attempt at finding an accurate sentiment classification approach (knowledge-based or machine learning) on Lithuanian Internet comments. Experiments will be performed with different pre-processing techniques, lexicons, and feature types.

3 The Lithuanian Language

In this section we discuss Lithuanian language properties focusing on those aspects (inflection morphology, word derivation system and word order in a sentence) that may be important in the sentiment classification task.

Lithuanian language has rich inflectional morphology, more complex than Latvian or Slavic languages (Savickienė et al., 2009). Adjectives are inflected by 7 cases, 2 (+1) genders, 2 numbers, 5 degrees of comparison, and have 2 pronominal forms; adverbs – by 5 degrees of comparison; nouns – by 7 cases, 2 (+1) genders and 2 numbers; verbs – by 3 moods, 4 tenses, 2 numbers, and 3 persons. Besides, verbs can have non-conjugative forms (participles, adverbial participles, verbal adverbs, and some forms of gerund) that can be inflected by tense, case, gender, number, and have an active or passive forms. Various inflection forms in Lithuanian language are expressed by the different endings (and suffixes), moreover, e.g. nouns have 12 different inflection paradigms; adjectives – 9.

Lithuanian language has rich word derivation system. 78 suffixes are used to derive diminutives and hypocoristic words (Ulvydas, 1965), that are especially frequent in spoken language; 25 prefixes are used for the nouns; 19 – for the verbs; and 3 (+4 in dialects) – for the adjectives and adjectival adverbs. Suffixes and prefixes change the meaning, e.g. suffix "-iaus-" change "geras" (*good*) to "geriausias" (*the best*) (by the way, the ending has to be adjusted to the new suffix, therefore "-as" is replaced by "-ias"); prefix "nu-" and reflexive participle "-si-" change "šnekėti" (*to talk*) to "nusišnekėti" (*to blunder out*). Prefixes in Lithuanian can also be used to derive phrasal verbs (e.g. from "eiti" (*to go*) to "įeiti" (*to go in*), "išeiti" (*to go out*), etc.) and negative words.

The particle "ne-" (*no*, *not*) or "nebe-" (*no longer*) giving to the words (adjectives, adjectival adverbs, adverbial adverbs, nouns, verbs and all their non-conjugative forms) an opposite meaning is attached to them as a prefix: "geras" (*good*) – "negeras" (*not good*); "skaisčiai" (*brightly*) – "nebeskaisčiai" (*no longer brightly*); "sėkmė" (*a fortune*) – "nesėkmė" (*a misfortune*); "bėgti" (*to run*) – "nebebėgti" (*no longer to run*); etc.

But if particle "ne", "nebe" or "nėra" (*no*, *not*) expresses contradiction, it is written separately (e.g. in "jis neblogas" (*he is not bad*) "ne" goes as the prefix, but in "jis ne blogas, o geras" (*he is not bad, but good*) "ne" goes separately.

The difference between English and Lithuanian is that a negative idea in English is expressed by only one negative word such as *nothing*, *nobody*, *never*, whereas in Lithuanian such sentence must contain two negated words, e.g. "niekas gerai nežaidžia" (*nobody plays well*) word-to-word translation is (*nobody well not plays*); "niekada nesakyk niekada" (*never say never*) word-to-word translation is (*never not say never*).

The word order in Lithuanian sentences is free, but it performs notional function, i.e. sentences are grammatically correct regardless of the word order, but the meaning (things that are highlighted) can differ. E.g. whereas in "tu esi labai geras" (*you are very good*) intensifier "labai" (*very*) is highlighted but in "tu esi geras labai" (*you are very good*) adjective "geras" (*good*) is highlighted, thus the first phrase gets higher positive intensiveness.

4 Methodology

4.1 Dataset

The dataset used in our sentiment classification task contains online Internet comments to articles crawled from the largest Lithuanian daily newspaper *Lietuvos rytas* (2013). These comments reflect people's opinions about the topical events in domestic and foreign politics, sport, etc.

All Internet comments were manually labeled as positive, negative or neutral. The decision about the class label was based on a mutual agreement of two human-experts. Efforts were made to focus solely on each comment, but known topic and previous posts could still influence experts' decision. Ambiguous comments were discarded thus leaving only single-labeled ones. Negative class strongly dominated the others. To maintain balanced class distribution the amount of comments (treated as instances in the classification process) belonging to the different classes was equalized by discarding redundant instances. See statistics of the dataset in Table 1.

Class label	**Number of instances**	**Number of tokens**	**Number of distinct tokens**
Positive	1,500	10,455	6,394
Negative	1,500	15,000	7,827
Neutral	1,500	13,165	4,039
Total	4,500	38,621	15,008

Table 1: Dataset statistics: the numbers were discarded; tokens (words) were transformed to lowercase.

The dataset contains texts representing informal Lithuanian language, i.e. texts are full of slang, foreign language insertions, and barbarisms. Besides, in the texts are a lot of typographical and grammatical errors. Moreover, Lithuanian language uses Latin script supplemented with diacritics, but in informal texts, diacritics (ą, č, ę, ė, į, š, ų, ū, ž) are very often replaced with matching Latin letters (a, c, e, e, i, s, u, u, z).

4.2 Classification methods

Sentiment classification task has never been solved for Lithuanian; therefore it is unclear which method could be the most suitable for the given dataset. Consequentially, in this research we will compare two different classification approaches – knowledge-based and machine learning – applying them on the informal texts.

The keystone of our knowledge-based approach is the lexicon that is applied to recognize sentiment words in the text. In our experiments we used two lexicons (see Table 2): manually labeled and automatically augmented one. Both lexicons are composed of 4 dictionaries: for adjectives, adverbs, nouns and verbs, respectively. Only lemmas (main words' forms containing ending and suffices/prefixes) are stored in the dictionaries.

The candidates for the first lexicon were extracted from 1 million running words taken from *Vytautas Magnus University Corpus* (Marcinkevičienė, 2000). These texts represent standard Lithuanian and were taken from six domains: fiction, legal texts, national newspapers, parliamentary transcripts, local newspapers, and popular periodicals. Words were transformed into their lemmas using Lithuanian part-of-speech tagger and lemmatizer *Lemuoklis* (Zinkevičius, 2000); (Daudaravičius et al., 2007) and transferred to the dictionaries containing appropriate parts-of-speech. Words in the first lexicon were manually labeled with their polarity values (-3/3 means that the word is strongly negative/positive; -2/2 – moderately negative/positive; -1/1 – weakly negative/positive; 0 – neutral). The decision was taken by mutual agreement of two human-experts that made efforts not to bind to the specific use cases, but consider only the most common sense of each word. The second lexicon was created by automatically augmenting the first one with the synonyms taken from *Lithuanian WordNet* (2013). Words from the manually labeled lexicon were used as the pre-selected "seeds" to search for the synonyms that automatically obtained the same polarity value and were added to the appropriate dictionaries.

Semantic orientation of each instance was determined by summing the polarity values of recognized sentiment words in the lemmatized texts. If total polarity value was positive (> 0), the instance was classified as positive; if negative (< 0) – as negative; if zero ($= 0$) – as neutral. E.g. "Filmas labai puikus" (*The film is great*) would be classified as positive, because *valueOf*("Filmas")=0 and *valueOf*("puikus")=3, thus $0 + 3 = 3 > 0$.

As the alternative for knowledge-based method we used two machine learning methods – i.e. Support Vector Machine (SVM), introduced by Cortes and Wapnik (1995) and Naïve Bayes Multinomial (NBM), introduced by Lewis and Gale (1994).

Polarity value	Adjectives	Adverbs	Verbs	Nouns	Total
-3	115 138	71 74	236 236	275 296	697 744
-2	151 175	120 122	333 337	719 775	1,323 1,409
-1	243 267	95 95	732 733	1,854 1,945	2,924 3,040
0	4,035 4,392	1,296 1,362	10,001 10,039	12,367 12,719	27,699 28,512
1	145 163	117 122	344 344	856 896	1,462 1,525
2	130 148	114 117	112 113	195 213	551 591
3	117 142	61 62	72 72	54 55	304 331
Total	4,936 5,425	1,874 1,954	11,830 11,874	16,320 16,899	

Table 2: Dictionaries statistics: the first value in each cell represents the number of items in manually labeled lexicon; the second – augmented with WordNet.

SVM is one of the most popular techniques for text classification, because it can cope with high dimensional feature spaces (e.g. 15,008 word features in our dataset); sparseness of feature vectors (e.g. among 15,008, each instance would have only ∼3.34 non-zero word feature values); and instances do not sharing any common features (common for short texts, e.g. average length of instance in our dataset is ∼8.58 words). Besides SVM does not perform aggressive feature selection which may result in a loss of information.

NBM method is also often used for text classification tasks (mostly due its simplicity): Naïve Bayes assumption of feature independence allows parameters of each feature to be learned separately. It performs especially well when the number of features is large. Besides, it is reported (e.g. by Pak and Parubek (2011)) that NBM can even outperform popular SVM in sentiment classification tasks.

In our experiments we used SMO kernel for SVM and NBM implementations in WEKA (Hall et al., 2009) machine learning toolkit, version 3.6[1]. All parameters were set to their default values.

4.3 Experimental setup

Before classification experiments tokens (i.e. words) in the dataset were pre-processed using different techniques. Knowledge-based method required lemmatization, whereas for machine learning methods lemmatization was optional. Despite that lemmatizer can solve disambiguation problems and achieve ∼0.94 accuracy on normative Lithuanian texts (Rimkutė and Daudaravičius, 2007); it could not recognize even ∼0.25 of words in our dataset.

Other optional pre-processing techniques involved emoticons replacement with appropriate sentiment words; Lithuanian diacritics replacements with appropriate Latin letters; and stop words removal.

Emoticons replacement demonstrated positive effect on English (Read, 2005) and triggered us to create such list for Lithuanian. The list contains 32 sentiment words (written in lemmas) with their appropriate and commonly used emoticon equivalents[2]. Thus, e.g. ":-)" would be replaced by "laimingas" (*happy*).

Words with replaced Lithuanian diacritics can neither be found in the dictionaries, nor recognized by the Lithuanian lemmatizer and therefore require special treatment. Whereas tools able to restore Lithuanian diacritics are not yet available, we have chosen opposite way by replacing all diacritics with matching Latin letters in the text, dictionaries and emoticons list and in such a way decreasing the number of unrecognized words (for knowledge-based method) and the sparseness of feature vector (for machine learning methods).

Stop words removal cannot affect the performance of knowledge-based method, but it can decrease the sparseness of the data for machine learning techniques. In our experiments we used stop words list with excluded interjections, because Spencer and Uchyigit (2012) showed that interjections are strong indicators of subjectivity.

Compulsory pre-processing steps included transformation of letters into lower-case, digits and punctuation removal. Statistics demonstrating the effect of different pre-processing techniques on the dataset are presented in Table 3.

Pre-processing was performed in such an order that previous steps could not harm following ones, thus lemmatization was performed before diacritics replacement, punctuation removal was performed after emoticons replacement, etc.

Knowledge-based method was evaluated using different combinations of dictionaries, whereas machine learning method – different types of features: *token unigrams* (the most common case);

[1] http://www.cs.waikato.ac.nz/ml/weka/.

[2] http://www.cool-smileys.com/text-emoticons.

token unigrams plus bigrams, i.e. token unigrams complemented with token bigrams (higher order n-grams sometimes outperform token unigrams); *token lemmas* (strongly recommended for highly-inflective languages); document-level *character 4-grams* (this type was reported as the best for Lithuanian topic classification by Kapočiūtė-Dzikienė et al. (2012)).

Class label	**Tokens after lemmatization**	**Tokens with emoticons**	**Tokens without stop-words**	**Tokens without diacritics**
Positive	10,386 3,177	10,664 4,027	8,982 3,941	10,455 3,724
Negative	14,928 6,475	15,107 7,811	11,945 7,716	15,000 7,457
Neutral	13,084 5,134	13,226 6,391	10,427 6,276	13,165 6,058
Total	38,398 11,669	38,997 14,966	31,354 14,923	38,621 13,983

Table 3: Pre-processed dataset statistics: the first value in each cell represents the number of all tokens, the second – distinct tokens. See Table 1 for unprocessed dataset statistics.

We expect the following statements to be confirmed experimentally: 1) emoticons replacement should increase the results since they usually reflect emotional state of the person; 2) diacritics replacement or lemmatization should improve the results by decreasing data sparseness and the number of unrecognized words; 3) all dictionaries should give better results for the knowledge-based method because contain more sentiment information; 4) machine learning methods should outperform knowledge-based approach because sentiments can be expressed in more complex ways.

5 Results

Accuracies (the number of correctly classified instances divided by all instances) of previously described experiments are summarized in Figure 1 – Figure 3.

Figure 1 summarizes the results obtained with the knowledge-based method. Figure 2 summarizes the results obtained with SVM method, Figure 3 – with NBM. 10-fold cross-validation was used in all experiments with machine learning methods.

6 Discussion

Since the balanced class distribution is maintained (see Table 1), both majority (probability to belong only to a major class) and random (the sum of squared probabilities of all classes) baselines are equal to 0.333. Figure 1 – Figure 3 show that obtained classification results are above the baseline.

The best results using knowledge-based method are achieved with *emoticons* and *diacritics replacement*, as expected (see Section 4.3), but emoticons replacement is more effective.

Augmented lexicon slightly outperforms manually labeled. Besides, *adjectives*, *nouns* and *verbs* improve the classification results for knowledge-based approach, but adverbs worsen it. Bad performance of adverbs contradicts our expectations. Analysis of erroneous cases revealed that very strong negative adverbs (used in slang) such as "baisiai" (*terribly*), "žiauriai" (*brutally*), etc. followed by the positive adjectives such as "geras" (*good*), "nuostabus" (*wonderful*) become positive intensifiers. Moreover, very often adverbs are found in the context does not expressing any sentiment at all, e.g. "gerai" (*well*) in "gerai pasakyta" (*well said*) should not be treated as positive word.

The results obtained with different machine learning methods – SVM and NBM are very contradictory and not always correspond to our expectations (see Section 4.3). In general the best feature type for SVM is either *token unigrams* or *token lemmas*; for NBM – *token unigrams plus bigrams*, but *token lemmas* is the second best result. Longer phrases (based on token bigrams) increase the sparseness of the data that seems to be harmful for SVM method, which does not perform aggressive feature selection. Whereas NBM is not as sensitive to it, *token unigrams plus bigrams* (carrying more sentiment information) give the best accuracy.

For both machine learning methods *token lemmas* are effective enough. The main problem is that Lithuanian lemmatizer could not recognize even a quarter of all words in the dataset, thus it can be assumed that this feature type could give even better results if lemmatizer would cope with informal Lithuanian language as well.

Results obtained by machine learning methods show that document-level *character 4-grams* (giving the best results for topic classification on Lithuanian texts) are not effective for sentiment classification. Character n-grams not only increase the sparseness, but result in a loss of important information about Lithuanian suffixes and prefixes. E.g. "gera" (*good*) and "negera" (*not*

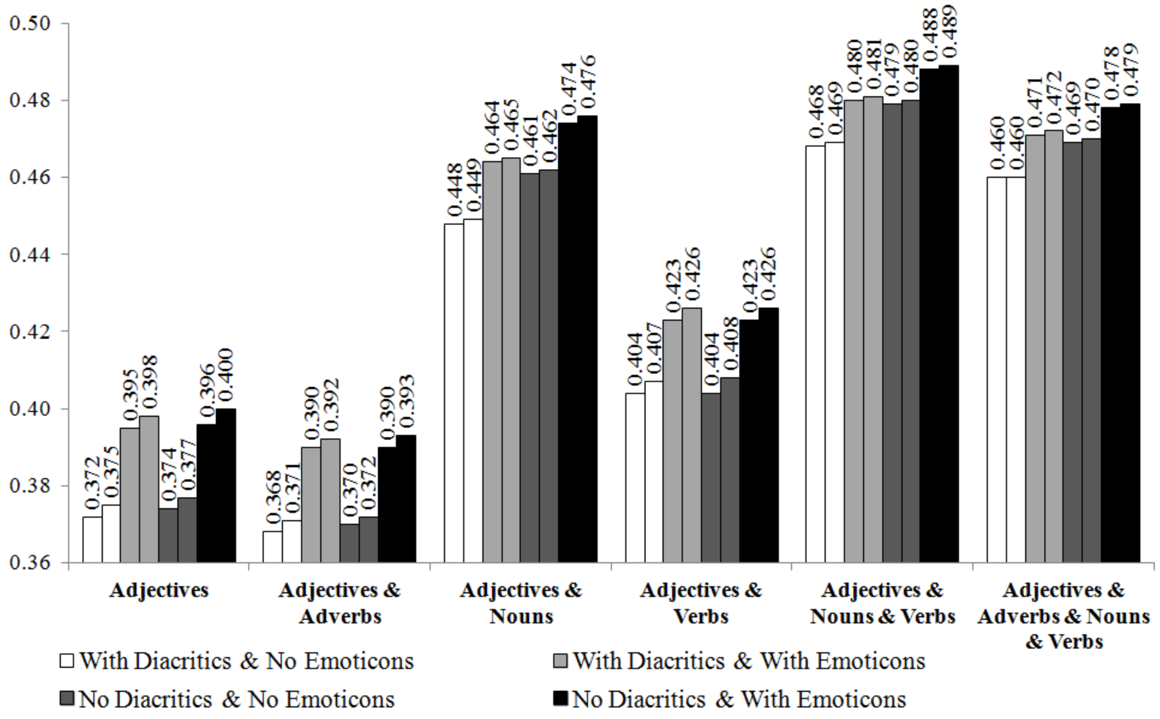

Figure 1: Accuracy of knowledge-based method, obtained using different lexicons and pre-processing techniques: groups of columns represent different combinations of dictionaries; shades of columns represent pre-processing techniques ("No Diacritics" stands for diacritics replacement, "With Diacritics" for no replacement, "With Emoticons" for emoticons replacement, "No Emoticons" for no replacement); the first column of the same shade represents results obtained using manually labeled lexicon, the second – augmented with WordNet.

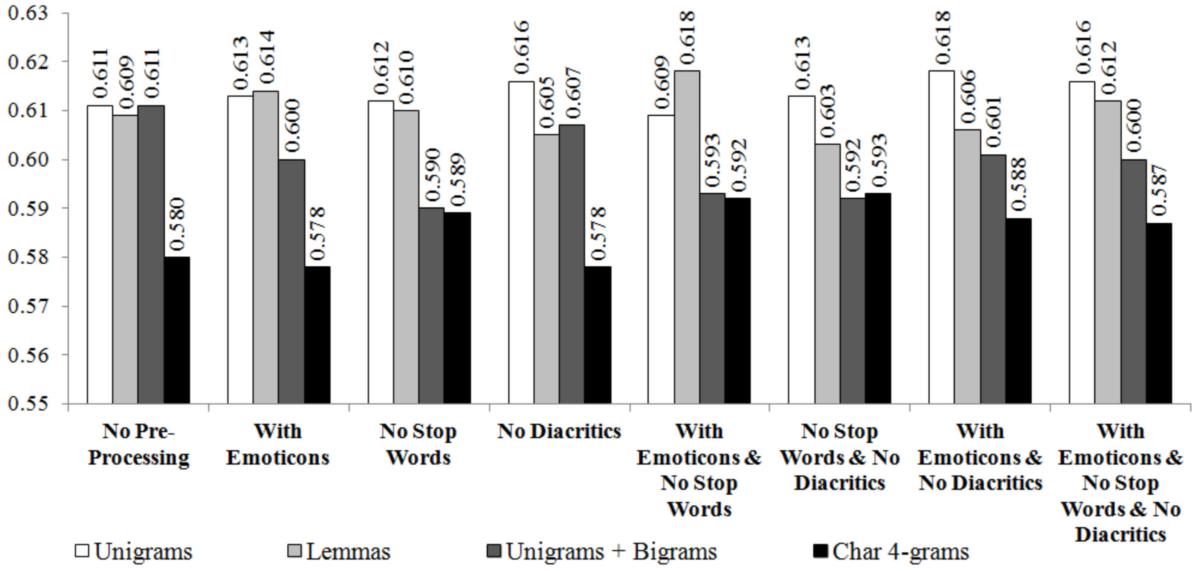

Figure 2: Accuracy of SVM method, obtained using different feature types and pre-processing techniques: groups of columns represent different pre-processing techniques ("With Emoticons" stands for emoticons replacement, "No Stop Words" for stop words removal, "No Diacritics" for diacritics replacement); shades of columns represent different feature types.

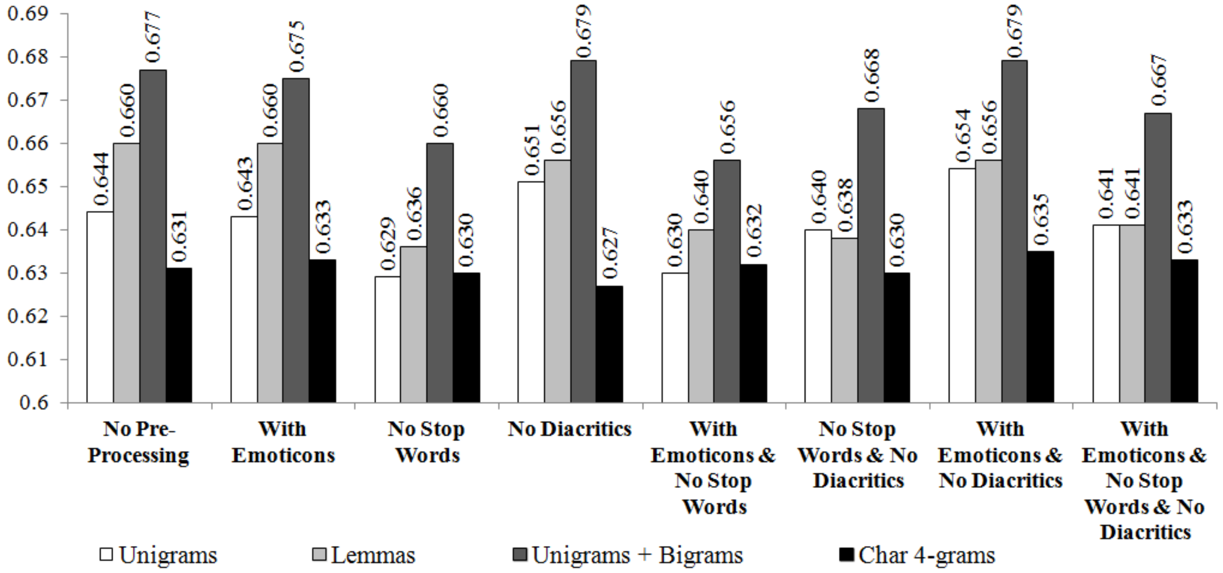

Figure 3: Accuracy of NBM, obtained using different feature types and pre-processing techniques.

good) contain the same 4 characters “gera”, but prefix “ne-” reverses the polarity.

As presented in Figure 2 and Figure 3 emoticons and diacritics replacement positively affect classification results, but the effect is much weaker compared to the knowledge-based approach. In general, for SVM there is no single pre-processing technique that could significantly stand out from the rest, while for NBM diacritics replacement is the best one, stop words removal is the worst. It can be assumed that despite stop words seem unimportant; they still carry sentiment information, especially significant using token bigrams.

As expected (see Section 4.3), machine learning methods significantly outperform knowledge-based. One of the main reasons is that the lexicons are not adjusted to a specific domain. Our goal was not to achieve as high accuracy as possible, but to determine a real potential of such method on informal Lithuanian texts. The analysis of erroneous cases revealed that adjectives, nouns and verbs are not the only sentiment indicators, e.g. interjection “valio!” (*hurray!*) in “valio! Auksas!” (*hurray! Gold!*) can express positive sentiment also.

Besides, diacritics replacement is still a considerable problem: e.g. whereas lexicon contains “šaunus” (*cool*, in masculine gender); the same word with replaced diacritics in feminine gender “sauni” will neither be recognized by lemmatizer, nor found in the lexicon with replaced diacritics.

The best result with knowledge-based method exceeds baseline by 0.156; with machine learning – by 0.346, but they are still low compared to the results obtained on English texts. Analysis of erroneous cases revealed that classifiers mostly fail due to the language variations when sentiments are expressed implicitly and require special treatment considering informal Lithuanian language specifics.

7 Conclusion and perspectives

In this paper we are solving Internet comments sentiment classification task for Lithuanian, using two different approaches: knowledge-based and machine learning.

Adjectives, nouns and verbs (excluding adverbs) are the most important sentiment indicators for the knowledge-based approach that was significantly outperformed by the machine learning methods. The best accuracy 0.679 is obtained using Naïve Bayes Multinomial with token unigrams plus bigrams as features and diacritics replacement as pre-processing technique.

In the future research we are planning to perform detailed class-wise error analysis that could help to find the solutions decreasing the number of erroneous cases. Besides, it would be interesting to experiment with the implicitly expressed sentiments.

Acknowledgments

This research is funded by European Union Structural Funds Project “Postdoctoral Fellowship Implementation in Lithuania” (VP1-3.1-ŠMM-01).

References

Anthony Aue and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and VS Subrahmanian. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of International Conference on Weblogs and Social Media (ICWSM)*.

Erik Boiy and Marie-Francine Moens. 2009. A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts. *Information Retrieval*, 12(5):526–558.

Henri Bouma, Olga Rajadell, Daniël Worm, Corné Versloot, and Harry Wedemeijer. 2012. On the early detection of threats in the real world based on open-source information on the internet. In *Proceedings of International Conference of Information Technologies and Security*.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *Proceedings of the International Confernece on Social Computing (SocialCom 2012)*, pages 71–80.

Alex Cheng and Oles Zhulyn. 2012. A System for Multilingual Sentiment Learning On Large Data Sets. In *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012)*, pages 577–592.

Yejin Choi and Claire Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 793–801.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20:273–297.

Hang Cui, Vibhu Mittal, and Mayur Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In *Proceedings of the Twenty First National Conference on Artificial Intelligence (AAAI-2006)*, pages 1265–1270.

Vidas Daudaravičius, Erika Rimkutė, and Andrius Utka. 2007. Morphological annotation of the Lithuanian corpus. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies (ACL'07)*, pages 94–99.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web (WWW'03)*, pages 519–528.

Xiaowen Ding and Bing Liu. 2007. The utility of linguistic rules in opinion mining. In *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 811–812.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, pages 417–422.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.

Tino Hartmann, Sebastian Klenk, Andre Burkovski, and Gunther Heidemann. 2011. Sentiment Detection with Character n-Grams. In *Proceedings of the Seventh International Conference on Data Mining (DMIN'11)*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'04)*, pages 168–177.

Matthew F. Hurst and Kamal Nigam. 2004. Retrieving topical sentiments from online document collections. In *Proceedings of Document Recognition and Retrieval*, volume XI, pages 27–34.

Jurgita Kapočiūtė-Dzikienė, Frederik Vaassen, Walter Daelemans, and Algis Krupavičius. 2012. Improving Topic Classification for Highly Inflective Languages. In *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1393–1410.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment classification of movie and product reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Ekaterina S. Kuznetsova, Natalia V. Loukachevitch, and Ilia I. Chetviorkin. 2013. Testing rules for a sentiment analysis system. In *Proceedings of International Conference Dialog*, pages 71–80.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR-94)*, pages 3–12.

Rūta Marcinkevičienė. 2000. Tekstynų lingvistika (teorija ir paktika) [Corpus linguistics (theory and practice)]. *Gudaitis, L. (ed.) Darbai ir dienos*, 24:7–63. (in Lithuanian).

Subhabrata Mukherjee and Pushpak Bhattacharyya. 2012. Sentiment Analysis in Twitter with Lightweight Discourse Analysis. In *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1847–1864.

Tony Mullen and Nigel Collier. 2004. Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 412–418.

Alexander Pak and Patrick Paroubek. 2011. Twitter for Sentiment Analysis: When Language Resources are Not Available. In *Proceedings of Database and Expert Systems Applications (DEXA 2011)*, pages 111–115.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundation and Trends in Information Retrieval*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 79–86.

Stephan Raaijmakers and Wessel Kraaij. 2008. Polarity Classification of Blog TREC 2008 Data with a Geodesic Kernel. In *Proceedings of the Seventeenth Text Retrieval Conference (TREC 2008)*, volume 500–277.

Jonathon Read and John Carroll. 2009. Weakly supervised techniques for domain-independent sentiment classification. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion (TSA'09)*, pages 45–52.

Jonathon Read. 2005. Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In *Proceedings of the 43th Annual Meeting on Association for Computational Linguistics (ACL'05) (Student Research Workshop)*, pages 43–48.

Erika Rimkutė and Vidas Daudaravičius. 2007. Morfologinis Dabartins lietuvių kalbos tekstyno anotavimas [Morphological annotation of the Lithuanian corpus]. *Kalbų studijos*, 11:30–35. (in Lithuanian).

Lietuvos Rytas. 2013. Lietuvos rytas. Internet daily newspaper, March. [http://www.lrytas.lt/] (in Lithuanian).

Ineta Savickienė, Vera Kempe, and Patricia J. Brooks. 2009. Acquisition of gender agreement in Lithuanian: exploring the effect of diminutive usage in an elicited production task. *Journal of Child Language*, 36(3):477–494.

James Spencer and Gulden Uchyigit. 2012. Sentimentor: Sentiment Analysis of Twitter Data. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 56–66.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. In *Proceedings of ACM Transactions on Information and System Security (TISSEC)*, pages 315–346.

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*, pages 417–424.

Kazys Ulvydas, editor. 1965. *Fonetika ir morfologija (daiktavardis, būdvardis, skaitvardis, įvardis) [Phonetics and morphology (noun, adjective, numeral, pronoun)]*, volume 1. Mintis, Vilnius, Lithuania. (in Lithuanian).

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. In *Proceedings of the 50th Annual Meeting on Association for Computational Linguistics (ACL'12)(System Demonstrations)*, pages 115–120.

Lithuanian WordNet. 2013. Lietuvių kalbos WordNet, February. [http://korpus.sk/ltskwn_lt.html] (in Lithuanian).

Vytautas Zinkevičius. 2000. Lemuoklis – morfologinei analizei [Morphological analysis with Lemuoklis]. *Gudaitis, L. (ed.) Darbai ir dienos*, 24:246–273. (in Lithuanian).