http://dx.doi.org/10.7592/FEJF2016.64.polarity

IDENTIFYING POLARITY IN DIFFERENT TEXT TYPES

Hille Pajupuu, Rene Altrov, Jaan Pajupuu

Abstract: While Sentiment Analysis aims to identify the writer's attitude toward individuals, events or topics, our aim is to predict the possible effect of a written text on the reader. For this purpose, we created an automatic identifier of the polarity of Estonian texts, which is independent of domain and of text type. Depending on the approach chosen – lexicon-based or machine learning – the identifier uses either a lexicon of words with a positive or negative connotation, or a text corpus where orthographic paragraphs have been annotated as positive, negative, neutral or mixed. Both approaches worked well, resulting in a nearly 75% accuracy on average. It was found that in some cases the results depend on the text type, notably, with sports texts the lexicon-based approach yielded a maximum accuracy of 80.3%, while over 88% was gained for opinion stories approached by machine learning.

Keywords: lexicon-based approach, machine learning approach, Naïve Bayes, polarity, sentiment analysis, SVM, text types

INTRODUCTION

The exponential increase in Internet usage and the diverse opportunities to express opinions on the Web have surrounded us with an unprecedented amount of information and different types of texts. To help us find a way through this abundance the recent decade has brought a boom in developing automatic text processing systems, which are able to extract, process and present relevant knowledge (Lloret et al. 2012). One of these systems is Sentiment Analysis (SA)[1], whose main goal is to determine whether a text, or a part of it, is subjective or not and, if subjective, whether it expresses a positive or negative view (Taboada 2016).

SA has mostly been performed for product reviews (incl. movie reviews, hotel reviews), forums, blogs, micro-blogs ("tweets"), news articles and social media in order to disclose the writer's opinions, attitudes and emotions toward individuals, events or topics (Pang & Lee 2008; Ravi & Ravi 2015).

SA has two main approaches to choose from: a lexicon-based approach and a machine learning approach.

The lexicon-based approach assumes that the text contains words with an emotional connotation, which are indicative of the writer's attitude (e.g., *abivalmis* 'helpful' is positive; *ebameeldiv* 'unpleasant' is negative). Analysis of the occurrence of such words in a text shows whether the text's polarity is rather positive or negative. Thus the lexicon-based approach requires a dictionary of words with emotional connotation, where the words carry a negative or positive label. The dictionary may also include information about connotation strength, that is, how positive or negative the word is. The target text is searched for dictionary words, which are respectively annotated in the text. A count of words with either polarity will lead to an assessment of the sentiment orientation of the text. In doing the counting, one may have to consider the possibility that the prior polarity of the word has been changed by context (Taboada 2016; Tabaoda et al. 2011; Wilson et al. 2009).

In the machine learning approach, the polarity of a text is determined by using classifiers. Classifiers are trained on a prepared corpus where documents or sentences have been annotated as either positive and negative, or as positive, negative and neutral. The most used classifiers in SA are Naïve Bayes and the Support Vector Machine (SVM). Classification has been attempted using different text features, such as considering unigrams only, bigrams, a combination of both, incorporating part-of-speech and position information, taking only adjectives, etc. Of all features, unigrams (individual words or tokens) have been the most effective. Accuracy has been increased by including all parts of speech, instead of being confined to adjectives, as well as by considering the position of the word in the text. SVM has generally given better results than Naïve Bayes, but Naïve Bayes also performs excellently, in particular if the feature space is small (Pang et al. 2002; Ravi & Ravi 2015; Taboada 2016; Vegda et al. 2014).

Both the lexicon-based and the machine-learning approach have yielded good results: for English texts, the accuracy is mostly between 70-90%, while higher scores have been achieved with texts sharing a domain (e.g., movie reviews), when subjected to a two-way classification (into positive and negative) (Pang & Lee 2008; Taboada et al. 2011; Zhang et al. 2014).

Which of the two approaches is preferable depends on the availability of sentiment resource, including annotated lexicons and corpora. The main problem with the lexicon-based approach lies in cross-domain adaptability. Words carrying a positive connotation in one domain may be negative or neutral in another. Thus, the lexicon has to be adapted to this or that domain. The machine-learning approach requires much human effort in document annotation and a good match between the training and testing data with respect to the domain (Taboada 2016; Zhang et al. 2014).

The resources are mainly available for English. Few other languages (e.g., Arabic, Spanish, Japanese, Chinese, French) have the necessary corpora or dictionaries. To bridge the gap, there have been attempts to translate some open source English resources into other languages by using the machine translation service. But every language has its peculiarities, which may render the adapting technique unsuitable, so one of the challenges of SA is to create specific linguistic resources for different languages (Balahur et al. 2014; Montoyo et al. 2012; Ravi & Ravi 2015; Taboada 2016; Zhang et al. 2014).

The approaches developed for SA can also be applied to other interlocking tasks.

While SA attempts to pinpoint the writer's attitude toward individuals, events or topics, our objective is to predict the possible effect of a written text on the reader. Unlike SA, where the main task is to tell facts apart from the writer's subjective opinion and analyse the latter, our underlying assumption is that any text, including an objective factual news text will affect the reader's mood, attitudes and decisions. Knowing the objective positive or negative polarity of the text would make this effect predictable.

The need to analyse text polarity ensues from the recent tendency to increasingly replace face-to-face interaction with the written one, both in private and working life. Unlike oral interaction, where the partner's mood and attitude can be detected from their voice and facial expression, written interaction leaves us face to face with nothing but a written text. Problems will arise if the text is misinterpreted. Whenever a writer is not sure what effect their text is likely to produce, an automatic identifier of text polarity could be of help in adjusting the text as necessary.

In addition to interaction, we are daily faced with a huge amount of texts, which makes deciding which ones to read difficult. Any means of automatic text processing to help us make the decision is welcome. One of the criteria motivating our decision is whether the text is positive or negative.

Another motive behind our wish to develop an automatic identifier of the polarity of a written text is a speech-technological need to make the result of text-to-speech synthesis more natural. The selection of the appropriate acoustic model for a text to be voiced requires information on the affectivity of the text (see Tamuri & Mihkla 2015).

Our challenge was to create an automatic identifier of text polarity for the Estonian language.

Estonian belongs to the Finnic branch of the Uralic language family. The language has a rich morphology, thus rendering it rather different from English, which is hitherto the dominating language in polarity identification. Morphological issues may, however, require special attention in polarity identification. For

Arabic, for example, which is also a morphologically rich language, the system includes automatic lemmatization (that is, head form retrieval) and part-of-speech (POS) tagging. This has yielded good results in subjectivity analysis (Abdul-Mageed et al. 2014).

For Estonian, the pilot study attempting to determine the polarity of a written text on the lexicon-based method was done manually on very limited material (Pajupuu et al. 2012a, 2012b). Inspired by the outcome we have set the task to create an automatic identifier of Estonian text polarity, which would be independent of domain and text type, and to find out whether the lexicon-based or machine learning method is more appropriate for the task.

RESOURCES NECESSARY FOR POLARITY IDENTIFICATION

First, a corpus had to be created to train and test the classifiers based on machine learning and to evaluate the accuracy of the lexicon-based method. According to Pajupuu and her colleagues (Pajupuu et al. 2012a, 2012b) it could be assumed that the optimal unit of polarity identification is an orthographic paragraph, not a full document or sentence. A paragraph is mostly a meaningful unit of text, mainly consisting of sentences of a similar polarity (see Figure 1).

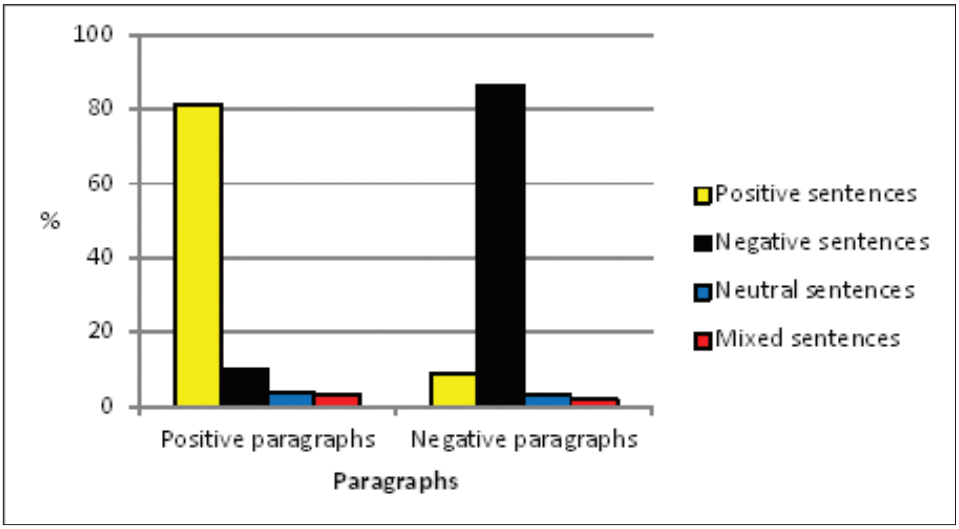

Figure 1. *According to the material of the Estonian Emotional Speech Corpus an orthographic paragraph usually consists of sentences of a similar polarity (Pajupuu et al. 2012b).*

The polarity corpus was compiled of articles of different rubrics of online dailies, weeklies, and reader comments, while the polarity of each paragraph was determined by native Estonian readers. Three subjects were asked to read the paragraphs independently of one another and decide from feeling whether the paragraph is positive, negative, neutral or ambivalent. The paragraphs were annotated using the dominant opinion (Pennebaker et al. 1997). The dominant opinion was the one expressed by at least two of the three readers. If no opinion dominated (all three were different, thus including, for example, positive, neutral and ambivalent), the paragraph was annotated as mixed (see Examples 1–4).

Example 1. A paragraph of an opinion story, annotated in corpus as positive:

Koht, mis varem ei olnud püha, võib selleks saada. Kui istutame tammikud, muudame need kohad pühaks. Hoolitseme ka selle eest, et tammikutes kasvaks kaunis kask ja püha pihlakas, et kaugete esivanemate vaimud end seal hästi tunneksid.[2]
A place that previously was not holy can become like that. We can make it holy ourselves by planting an oak forest. Moreover, let us take care that the oak forest also features the beautiful birch and the protective rowan, just to make the distant ancestral spirits feel good.

Example 2. A paragraph of a crime news story, annotated in corpus as negative:

Tabati ka üks kriminaalses joobes sõidukijuht. See juhtus pühapäeva öösel kella 4 ajal, kui Viljandis Lääne tänaval peeti kinni sõiduauto BMW, mille roolis oli 21-aastane noormees. Tema suhtes alustati kriminaalmenetlust.[3]
Also, a criminally intoxicated driver was apprehended. It happened at 4 o'clock Sunday morning that a BMW driven by a 21-year old was stopped on Lääne St. in Viljandi. Criminal charges were filed.

Example 3. A paragraph of culture news, annotated in corpus as neutral:

Peaaegu samasugune nägi pööning välja märtsis, kui kunstnik oli sinna üles seadnud "Asjade" esimese osa. Vahepealse kuue kuu jooksul on katusealune ja seda külastanud vaatajad osa saanud suurtest muudatustest.[4]
The attic looked almost the same in March, just after the artist had set up the first part of the "Things". During the six months passed, the attic and its visitors have been exposed to some considerable changes.

Example 4. A paragraph of domestic news, annotated in corpus as mixed:

Uuringust tuli välja, et ligi pooled inimesed ei kavatse enam Eestisse tagasi tulla, kuid paljud vastajad tunnistasid, et kui Eestis oleks neil rohkem väljakutseid ja huvitav töö, siis kaaluksid nad tagasitulemist.[5] According to the results, nearly half of the people had no intention of returning to Estonia, however, many respondents admitted that if Estonia offered them more challenges and an interesting job, they would reconsider.

In total, the corpus contains 4,086 annotated paragraphs, see Figure 2 and Table 1. The corpus is an open source and available for free.[6]

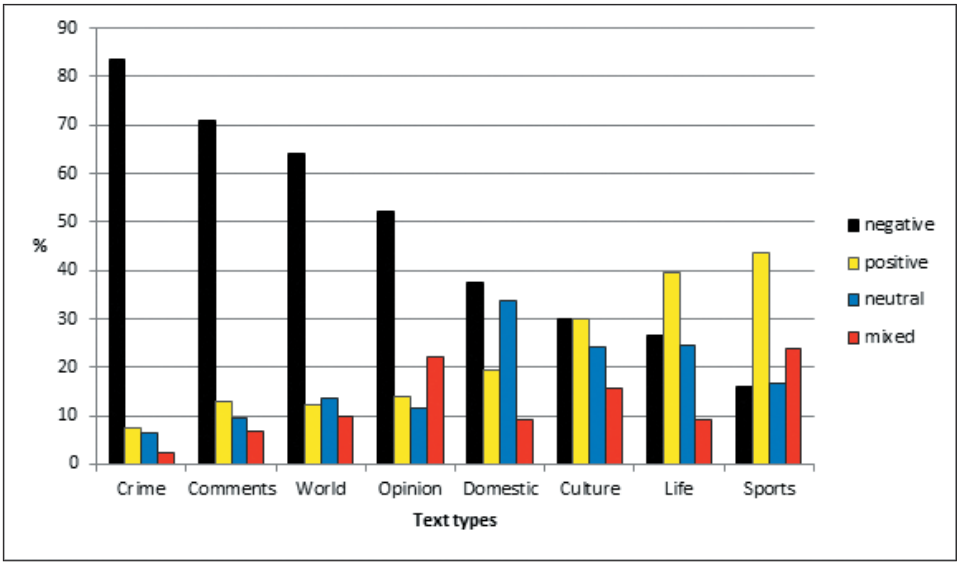

Figure 2. *The annotated corpus paragraphs by text types.*

The lexicon to be used in polarity identification was compiled of words with a positive or negative connotation. We dropped the idea of translating relevant dictionaries from other languages, because every culture has specific words carrying a positive or negative connotation precisely for the members of this particular culture. For a local Estonian person, for example, the positive words

include *leib* 'bread', *vaikne* 'quiet', *sõltumatu* 'independent', whereas *hilinema* 'be late' and *vihmane* 'rainy' are negative. In some other culture the same words may be neutral or even of an opposite connotation.

The volume of the dictionaries used in lexicon-based SA can be very different ranging from the 5,000 polarity-annotated words as in SO-CAL (Taboada et al. 2011) to the nearly 76,000 words of the Macquarie Semantic Orientation Lexicon (Mohammad et al. 2009). The optimal size is still open to discussion. Taboada and her colleagues have found that a large dictionary tends to capture more noise, leading to inaccurate results in SA (Taboada 2016; Taboada et al. 2011). Pajupuu and her colleagues (Pajupuu et al. 2012b) have observed that a relatively small dictionary of frequent words can turn out to be efficient, because most of the frequent polar words are monovalent, so that their connotation is seldom changed by context (e.g., the frequent Estonian word *koostöö* 'cooperation' is invariably positive, whereas the relatively rare word *vähenõudlik* 'undemanding' can be either positive or negative depending on the context).

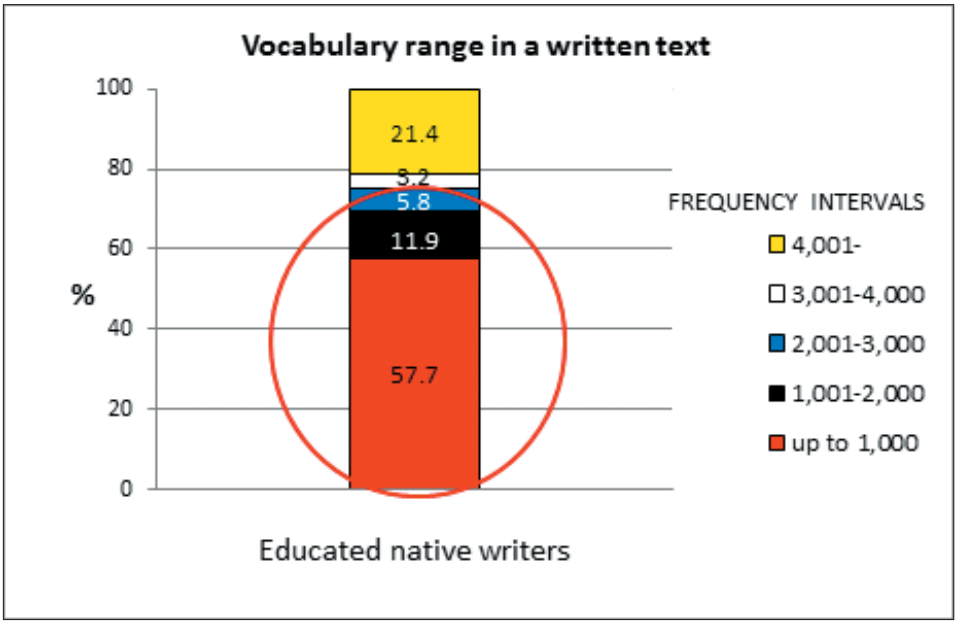

Figure 3. *Vocabulary range in demanding texts by educated native writers (Kerge et al. 2014).*

As our aim was to identify the polarity of texts regardless of their domain and text type, we decided to limit the dictionary to frequent words for the reason that frequent words tend to appear in most texts, whereas rare words rather belong to specific topics or domains. In demanding texts produced by educated writers most (75%) of the words belong to the 3,000 more frequent ones (Kerge et al. 2014), see Figure 3. Hence, if a text has an emotional meaning, it is likely to contain some frequent words with a polarity.

Another question deriving from SA practice is whether the dictionary should only consist of adjectives or should other parts of speech be included (see Taboada 2016). We decided to determine the positive or negative connotation, if any, for all words included in the Basic Estonian Dictionary regardless of their part of speech (see Vainik 2012). The Basic Estonian Dictionary contains the 3,015 most frequent Estonian words. Like in the case of orthographic paragraphs we asked four native speakers of Estonian to determine whether the words have a positive or negative connotation, are neutral or ambivalent (in the latter case polarity depends on context). Due to the dominating opinion, words with a positive (317) or negative (322) connotation were included in the polarity dictionary (cf. Pajupuu et al. 2012b). After supplementing the list with some antonyms and derivatives the polarity dictionary now features 617 words with a positive and 730 words with a negative annotation.

As Estonian is a language of considerable morphological richness and text words can manifest very different grammatical forms, we first developed an idea to use a lemmatizer to reduce all different forms of a word to the dictionary head-word (cf. Abdul-Mageed et al. 2014). The idea was soon dropped, however, because different forms of a word can be of different polarities (e.g., *abi* (pos) 'help' *abita* (neg) 'without help'; *nautima* (pos) 'enjoy', *nautimata* (neg) 'without enjoying'). So we decided to include the words with all their grammatical forms and annotate them as positive or negative. The final size of the dictionary was 38,628 tokens. The addition of grammatical forms revealed cases of morphological homonymy (accidental coincidence between certain grammatical forms of words of different parts of speech) (e.g., *tänavat* (pos) '(he) is known to say thank you' and *tänavat* (neutr) 'this street PartSg'; *lood* (pos) 'you'll create (sth)', *lood* (neutr) 'stories'; *mees* (pos) 'in honey' *mees* (neutr) 'a man'; *mürgita* (pos) 'without poison' and *mürgita* (neg) 'just poison (him)'. Such homonymous forms were excluded from the dictionary.

Next, we used the paragraph corpus to see how many paragraphs contain words included in our dictionary. As the paragraph corpus contains not only positive, negative and mixed paragraphs, but also neutral ones, the results can be regarded as promising, see Table 1.

Table 1. *Paragraph corpus and the number of paragraphs containing at least one word included in the polarity dictionary.*

Text types	**Number of paragraphs**	**Number of paragraphs with a word in the polarity dictionary**	
Opinion	972	876	90.1%
Domestic	419	287	68.5%
Life	518	381	76.3%
Comments	1008	781	77.5%
Crime	209	162	77.5%
Culture	262	183	69.8%
Sports	385	311	80.8%
World	313	239	76.4%

METHOD

In the lexicon-based approach, each text word was compared with those in the polarity dictionary. The following rules were applied:
a) A negation (*ei* or *ega*) will make the following positive word negative, for example, *rõõmustama* (pos), 'to please; rejoice', *ei rõõmusta* (neg) 'will not be/make happy';
b) A negation will make the following negative word positive, for example, *valetama* (neg) 'to lie', *ei valeta* (pos) 'does not lie';
c) A negation will make the following neutral word negative, for example, *tulema* (neutr) 'to come', *ei tule* (neg) 'is not coming'.

Then the positive and negative words were summarized over each paragraph. The paragraph was classified as positive or negative according to the connotation of the dominating words. For example, if a paragraph contained just one positive word (+1) and four negative ones (-4), the paragraph was classified as negative. If the number of positive and negative words was equal, the paragraph was classified as mixed. If there were neither any positive or negative words, the paragraph was classified as neutral (cf. Pajupuu et al. 2012b).

The lexicon-based method was tested on the whole material of the paragraph corpus.

In machine learning the words (unigrams) of the paragraphs were used as features. Two classifiers, Naïve Bayes and the Support Vector Machine (SVM) were tested in the Scikit-learn environment (Pedregosa et al. 2011).

Naïve Bayes was trained using:

```
Pipeline([
    ('vect',CountVectorizer(binary=True)),
    ('clf', MultinomialNB())
    ])
```

SVM was trained using:

```
    Pipeline([
        ('tfidf', TfidfVectorizer(use_idf=True)),
        ('clf', SGDClassifier(loss='squared_hinge', penalty='l2', random_state=None,
        alpha=1e-3))
    ])
```

Both Naïve Bayes and the SVM were trained all over the corpus without discriminating between text types, while a hundred paragraphs of each text type were used for testing. The paragraphs annotated as mixed in corpus were removed from the training material, because in pilot tests such paragraphs were found to have a considerable lowering effect on the accuracy of identification. A three-way classification (positive, negative, neutral) was used.

In order to evaluate the polarity classification accuracy of the lexicon-based method versus machine learning their outputs were compared with the human ratings of the corpus paragraphs. Automatic classification was considered correct if it coincided with the human one, plus the positive–mixed and negative–mixed pairs.[7]

From the polarity of the orthographic paragraphs found either by the lexicon-based or machine learning method, the polarity of the full document was determined. First, the words in a paragraph were counted. Each word was assigned the polarity of the paragraph it belonged to, that is, if a paragraph was positive (negative, neutral, mixed), all its words got a respective positive (negative, neutral, mixed) label.

If a document consisted of paragraphs of one and the same class, the document was assigned the same polarity[8]: POSITIVE, NEGATIVE, NEUTRAL or MIXED.

If a document contained paragraphs of two classes, the proportion of the words of either class in the document was calculated. If the words of a class (positive, negative, neutral, or mixed) made up at least 66.7% of the document, the latter was respectively labelled as MOSTLY POSITIVE, MOSTLY NEGATIVE, MOSTLY NEUTRAL or MOSTLY MIXED. If the words of none of the classes reached 66.7% of the document words, the document was labelled as MOSTLY MIXED.

If a document contained paragraphs of three or four classes, the document was labelled after the class whose words made up at least half of the document words (MOSTLY POSITIVE, MOSTLY NEGATIVE, MOSTLY NEUTRAL or MOSTLY MIXED). In the rest of cases the document was labelled as MOSTLY MIXED.

The accuracy of the labelling of full documents was not evaluated, because the corpus does not include full documents with human-determined or human-annotated polarity.

RESULTS AND DISCUSSION

Our aim was to create an automatic identifier of Estonian text polarity, which would be independent of domain and text type, and find out whether the lexicon-based or machine learning approach should be preferred in doing so.

Table 2 presents, by text types, the percentage of paragraphs with accurately identified polarity as found by using the lexicon-based approach versus the machine learning method using the classifiers Naïve Bayes and SVM.

Table 2. *Percentage of paragraphs with accurately identified polarity, by text types.*

Text types	**Number of paragraphs**	**Lexicon-based approach**	**Machine learning approach**			
			Naïve Bayes		**SVM**	
			M	SD	M	SD
Opinion	972	76.4	88.8	1.8	88.2	1.3
Domestic	419	73.0	62.7	1.7	67.2	1.8
Life	518	62.4	64.5	2.7	69.5	2.4
Comments	1008	58.3	86.0	2.9	85.2	1.4
Crime	209	78.0	87.9	2.0	87.7	1.2
Culture	262	78.2	54.4	5.2	58.1	2.3
Sports	385	80.3	75.7	3.0	75.2	2.6
World	313	79.2	68.7	1.1	72.5	2.5
M		*73.2*	*73.6*		*75.5*	
SD		*8.3*	*13.0*		*10.8*	

Note. *For Naïve Bayes and SVM an average of five tests is presented.*

In the lexicon-based approach the paragraphs were classified into four polarity classes (positive, negative, neutral, mixed). The accuracy was between 58.3–80.3, being lowest for comments and highest for sports articles. In the machine

learning the paragraphs were classified into three polarity classes (positive, negative, neutral). Naïve Bayes yielded accuracy readings from 54.4–88.8, the lowest score belonging to cultural texts and the highest to opinion stories. With SVM the accuracy ranged from 58.1–88.2, being also lowest for cultural texts, but the highest accuracy was scored by opinion stories. There was no substantial difference between the two approaches in polarity identification (the average score being 73.2–75.5), but the accuracy scores did differ across text types.

In the lexicon-based approach, the lowest accuracy was measured in the case of comments. One of the reasons is probably that comments are often very short, consisting of a single sentence or just a word, whereas the optimal unit for lexicon-based identification of polarity is an orthographic paragraph, that is, at least two semantically connected sentences. A single sentence, however, may not contain any words from the polarity dictionary and so the sentence will be classified as neutral. For example, the sentence *Poodi on tarvis, aga ehitage kellegi teise kodu kõrvale* (A store is necessary, but build it next to someone else's home) was classified as negative by humans, but neutral by the lexicon-based identifier. Moreover, comments often feature unusual word usage (e.g., swear words, abbreviations, slang, foreign words and expressions) and deviation from regular orthography. If a word is relatively rare and has deviant orthography, it cannot be found in our polarity dictionary and will consequently be regarded as neutral (e.g., the two-word sentence *krdi ajukääbik* (you damned fool), which consists of an abbreviation and a rare derogatory word, was classified as negative by humans, but neutral by the lexicon-based identifier). The machine learning method, however, copes well with comments, gaining an accuracy of 86.0% with Naïve Bayes and 85.2% with SVM.

In machine learning the polarity identification accuracy was the lowest with paragraphs about culture. Further analysis should disclose whether the reason could lie in the small number of culture paragraphs in the training corpus, and in their lexical diversity. Actually, crime, with a still smaller number of paragraphs and low lexical variation, ended up with a very high accuracy indeed.

Although our aim was to find out which approach, the lexicon-based one or machine learning, works better for the identification of polarity in Estonian texts, the answer is still ambiguous. According to the mean accuracy (~75%) both approaches can be regarded as equally appropriate and worthy of further development. The lexicon-based approach requires the existence of a polarity dictionary, while machine learning requires a corpus of polarity-annotated texts (see e.g., Balahur et al. 2014; Taboada 2016). Both are now available for the Estonian language, open for public use, improvement and extension.

The results of polarity identification for Estonian written texts are not quite comparable with the SA results available for some other languages. SA has a different purpose, notably, to identify the writer's attitude towards an entity or topic (see e.g., Montoyo et al. 2012). Our aim was to identify the polarity of a text in order to predict its possible effect on the reader. Any text, however subjective or objective, can carry a positive or negative meaning (cf. Patel et al. 2015). On the one hand, the task of SA is somewhat more complicated in that first, one has to discern and separate the subjective part of the text and then identify the polarity of this part, leaving the objective part aside. On the other hand, however, SA looks rather more simple, because it is mostly domain and text-type centred and in many cases the classification used is dichotomous, dividing its objects into positive and negative ones (see Ravi & Ravi 2015). Our identifier is neither domain or text-type centred and the objects are divided either between four classes (as in the lexicon-based approach) or between three classes (as in machine learning). Thus, for us the 70–90% accuracy rates of SA prevalently scored on English material can be regarded as an approximate benchmark only. Our mean accuracy of ~75% is two to three times better than chance probability, which can be considered a sufficiently good score for a polarity identifier that is independent both of text type and of domain.

Little is known of SA studies for languages with a rich morphology. For SA performed on Arabic social media texts a lemmatizer and POS-tagging were used and their dichotomous classification into positive and negative yielded accuracies of 70.3–81.8%, depending on the text type (Abdul-Mageed et al. 2014). Our polarity identifier, using a three-way and a four-way classification, gave a mean accuracy of the same interval. Whether lemmatization and POS-tagging could raise the accuracy even more remains to be tested. Our accuracy currently achieved in polarity identification by the machine learning method is consistent with the results used to prove that the simple use of unigrams as features leads to good results (cf. Balahur et al. 2014; Pang & Lee 2008; Vegda et al. 2014).

Our results from the lexicon-based approach corroborate the statement of Taboada et al. (2011) that small dictionaries can do very well for SA. Their nearly 5,000-word dictionary worked better in their Semantic Orientation CALculator (SO-CAL) (mean accuracy 78.7%) than bigger dictionaries. Our polarity identifier, using our polarity dictionary of 1,347 frequent words with a positive or negative connotation performed similarly well. As far as we know, our dictionary is one of the smallest used in this field. As the compilation of a polarity dictionary of frequent words of this amount is a relatively simple task requiring little human resources, whatever the language, it is well worth giving it a try.

Our success is largely due to the introduction of the orthographic paragraph as a unit of identification. If a person labels a paragraph as negative or positive, the paragraph usually contains some frequent words of a negative or positive connotation, respectively, enabling our lexicon-based polarity identifier to perform, in several text domains, rather similarly to humans.

The limitation of the present study is the relatively small number of domains and text types involved. There may still be domains and text types requiring an upgrade of both the training corpus and the dictionary. The necessity is implied by the relatively lower accuracy of comments polarity identification in the case of the lexicon-based approach. The written texts whose style resembles that of oral speech (real time messages, blog, tweets, comments etc.) certainly need the polarity dictionary to be supplemented with frequent polar words characteristic of their text types.

By way of conclusion, we have created resources for polarity identification in the Estonian language (one of the Finnic branch of the Uralic language family) and tested both the lexicon-based and the machine learning approaches to the classification of texts by polarity[9]. The results look promising and the problems revealed are worth further investigation, including in the SA direction.

ACKNOWLEDGEMENTS

This work was supported by institutional research funding IUT 35-1 of the Estonian Ministry of Education and Research, TK 145 “Centre of Excellence in Estonian Studies – CEES”, and the governmental basic financing of the Institute of the Estonian Language from the Estonian Ministry of Education and Research.

NOTES

[1] Sentiment Analysis, Opinion Mining, and Subjectivity Analysis are broadly used as synonyms, although some sources make a difference between the concepts (Pang & Lee 2008; Serrano-Guerrero et al. 2015).

[2] Quoted from Sutrop, Urmas. Nagu Taara tammikud. [Like Taara oak-woods.] Newspaper *Maaleht*, December 19, 2014. Available at http://maaleht.delfi.ee/news/maaleht/arvamus/urmas-sutrop-nagu-taara-tammikud?id=70389341, last accessed on May 20, 2016.

[3] Quoted from Teder, Merike. Tabatud joobes juhid saadeti arestimajja. [Caught drunk drivers were sent to detention.] Newspaper *Postimees*, September 11, 2012. Available at http://www.postimees.ee/968740/tabatud-joobes-juhid-saadeti-arestimajja, last accessed on May 19, 2016.

[4] Quoted from Hanson, Raimu. Pööningu kilast-kolast käis üle jumalik hingus. [Divine breath flew over the junk in the attic.] Newspaper *Tartu Postimees*, September 13, 2012. Available at http://tartu.postimees.ee/970946/pooningu-kilast-kolast-kais-ule-jumalik-hingus, last accessed on May 19, 2016.

[5] Quoted from Traks, Kristina. Mis tooks Eesti töötajad Soomest tagasi? [What would bring Estonian workers back from Finland?] Newspaper *Postimees*, September 13, 2012. Available at http://majandus24.postimees.ee/971422/mis-tooks-eesti-tootajad-soomest-tagasi, last accessed on May 19, 2016.

[6] Free and open corpus of paragraphs http://peeter.eki.ee:5000/valence/paragraphsquery/.

[7] An orthographic corpus paragraph has been annotated as mixed in two cases: (1) the readers have determined the paragraph as ambivalent, which means that it contains both positive and negative elements; (2) there is no dominant opinion (e.g., half of the readers have determined the paragraph as negative, the other half as positive). Therefore, we decided that a "mixed" paragraph can be considered correctly identified by the program if it classifies the paragraph as positive or negative.

[8] Polarity in a wider sense includes four classes: positive, negative, neutral and mixed.

[9] Automatic identifier of written text polarity, http://peeter.eki.ee:5000/valence/, and https://github.com/EKT1/valence/.

REFERENCES

Abdul-Mageed, Muhammad & Diab, Mona & Kübler, Sandra 2014. SAMAR: Subjectivity and Sentiment Analysis for Arabic Social Media. *Computer Speech and Language*, Vol. 28, No. 1, pp. 20–37. http://dx.doi.org/10.1016/j.csl.2013.03.001.

Balahur, Alexandra & Mihalcea, Rada & Montoyo, Andres 2014. Computational Approaches to Subjectivity and Sentiment Analysis: Present and Envisaged Methods and Applications. *Computer Speech and Language*, Vol. 28, No. 1, pp. 1–6. http://dx.doi.org/10.1016/j.csl.2013.09.003.

Kerge, Krista & Pajupuu, Hille & Alp, Pilvi & Põlda, Halliki & Uusen, Anne 2014. Towards Sophisticated Writing. *Proceedings of the Tallinn University Institute of Estonian Language and Culture: Studies in Language Acquisition, Learning, and Corpora*, Vol. 16, pp. 103–115. Available at http://www.academia.edu/9729151/Towards_sophisticated_writing, last accessed on May 19, 2016.

Lloret, Elena & Balahur, Alexandra & Gómez, José & Montoyo, Andrés & Palomar, Manuel 2012. Towards a Unified Framework for Opinion Retrieval, Mining and Summarization. *Journal of Intelligent Information Systems*, Vol. 39, No. 3, pp. 711–747. http://dx.doi.org/10.1007/s10844-012-0209-4.

Mohammad, Saif & Dorr, Bonnie & Dunne, Cody 2009. Generating High-Coverage Semantic Orientation Lexicons from Overtly Marked Words and a Thesaurus. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-2009)*, August 6–7, Singapore, pp. 599–608.

Montoyo, Andrés & Martínez-Barco, Patricio & Balahur, Alexandra 2012. Subjectivity and Sentiment Analysis: An Overview of the Current State of the Area and Envisaged Developments. *Decision Support Systems*, Vol. 53, No. 4, pp. 675–679. http://dx.doi.org/10.1016/j.dss.2012.05.022.

Pajupuu, Hille & Kerge, Krista & Altrov, Rene 2012a. Detecting Emotional Valence of Text by Using a Small Dictionary. In: Izaskun Elorza & Ovidi Carbonell i Cortés & Reyes Albarrán & Blanca García Riaza & Miriam Pérez-Veneros (eds.) *Empiricism and Analytical Tools for 21st Century Applied Linguistics. Selected Papers from the XXIX International Conference of the Spanish Association of Applied Linguistics (AESLA)*, Colección Aquilafuente 185. Salamanca: Universidad de Salamanca, pp. 229–242.

Pajupuu, Hille & Kerge, Krista & Altrov, Rene 2012b. Lexicon-Based Detection of Emotion in Different Types of Texts: Preliminary Remarks. *Eesti Rakenduslingvistika Ühingu aastaraamat / Estonian Papers in Applied Linguistics*, Vol. 8, pp. 171–184. DOI:10.5128/ERYa8.11.

Pang, Bo & Lee, Lillian 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, Vol. 2, Nos. 1–2, pp. 1–135. http://dx.doi.org/10.1561/1500000011.

Pang, Bo & Lee, Lillian & Vaithyanathan, Shivakumar 2002. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vol. 10, pp. 79–86. DOI:10.3115/1118693.1118704.

Patel, Vishakha & Prabhu, Gayatri & Bhowmick, Kiran 2015. A Survey of Opinion Mining and Sentiment Analysis. *International Journal of Computer Applications*, Vol. 131, No. 1, pp. 24–27. http://dx.doi.org/10.5120/ijca2015907218.

Pedregosa, Fabian & Varoquaux, Gaël & Gramfort, Alexandre & Michel, Vincent & Thirion, Bertrand & Grisel, Olivier & Blondel, Mathieu & Prettenhofer, Peter & Weiss, Ron & Dubourg, Vincent & Vanderplas, Jake & Passos, Alexandre & Cournapeau, David & Brucher, Matthieu & Perrot, Matthieu & Duchesnay, Édouard 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830. Available at http://www.jmlr.org/papers/v12/pedregosa11a.html, last accessed on May 19, 2016.

Pennebaker, James W. & Mayne, Tracy J. & Francis, Martha E. 1997. Linguistic Predictors of Adaptive Bereavement. *Journal of Personality and Social Psychology*, Vol. 72, No. 4, pp. 863–871. http://dx.doi.org/10.1037/0022-3514.72.4.863.

Ravi, Kumar & Ravi, Vadlamani 2015. A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications. *Knowledge-Based Systems*, Vol. 89, pp. 14–46. http://dx.doi.org/10.1016/j.knosys.2015.06.015.

Serrano-Guerrero, Jesus & Olivas, Jose A. & Romero, Francisco P. & Herrera-Viedma, Enrique 2015. Sentiment Analysis: A Review and Comparative Analysis of Web Services. *Information Sciences*, Vol. 311, pp. 18–38. DOI:10.1016/j.ins.2015.03.040.

Taboada, Maite 2016. Sentiment Analysis: An Overview from Linguistics. *Annual Review of Linguistics*, Vol. 2, pp. 325–347. http://dx.doi.org/10.1146/annurev-linguistics-011415-040518.

Taboada, Maite & Brooke, Julian & Tofiloski, Milan & Voll, Kimberly & Stede, Manfred 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, Vol. 37, No. 2, pp. 267–307. http://dx.doi.org/10.1162/COLI_a_00049.

Tamuri, Kairi & Mihkla, Meelis 2015. Expression of Basic Emotions in Estonian Parametric Text-To-Speech Synthesis. *Eesti ja soome-ugri keeleteaduse ajakiri / Journal of Estonian and Finno-Ugric Linguistics*, Vol. 6, No. 3, pp. 145–168. http://dx.doi.org/10.12697/jeful.2015.6.3.06.

Vainik, Ene 2012. Kuidas määrata eesti keele sõnavara tundetoone? [Detecting Emotional Valencies for the Estonian Vocabulary.] *Eesti Rakenduslingvistika Ühingu aastaraamat / Estonian Papers in Applied Linguistics*, Vol. 8, pp. 257–274. DOI: http://dx.doi.org/10.5128/ERYa8.17.

Vegda, Hiteshkumar N. & Patel, Tejal R. & Patel, Bhargesh B. 2014. A Survey on Sentiment Analysis of Texual (sic) Review. *International Journal of Research in Advent Technology*, Vol. 2, No. 2, pp. 2321–9637. Available at http://www.academia.edu/9310690/A_survey_on_Sentiment_Analysis_of_Texual_Review, last accessed on May 19, 2016.

Wilson, Theresa & Wiebe, Janyce & Hoffmann, Paul 2009. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics*, Vol. 35, No. 3, pp. 399–433. http://dx.doi.org/10.1162/coli.08-012-R1-06-90.

Zhang, Hailong & Gan, Wenyan & Jiang, Bo 2014. Machine Learning and Lexicon Based Methods for Sentiment Classification: A Survey. *11th Web Information System and Application Conference*, 12–14 September 2014, Tianjin, China, pp. 262–265. DOI: 10.1109/WISA.2014.55.